Modelo para predicción de potencia de paneles fotovoltaicos utilizando técnicas de clasificación no supervisada y redes neuronales artificiales.

Johanna Michelle Romero Rodríguez

Universidad del Norte

Departamento d ingeniería mecánica

Barranquilla, Colombia

2020

Modelo para predicción de potencia de paneles fotovoltaicos utilizando técnicas de clasificación no supervisada y redes neuronales artificiales.

Johanna Michelle Romero Rodríguez

Tesis entregada como requisito para optar por el título de:

Magíster en ingeniería mecánica

Director:

Antonio José Bula Silvera PhD.

Grupo de investigación:

Uso Racional de la Energía y Medio Ambiente: UREMA

Universidad del Norte

Departamento de ingeniería mecánica

Barranquilla, Colombia

2020

A יהוה,

La roca de mi salvación, quien siempre me ha sostenido.

Agradecimientos

Al YHVH, que, siendo el creador y rey del universo, camina de mi lado. A él doy gracias por darme todo lo que tengo.

A mis padres, Néstor y Mónica, que siempre estuvieron ayudándome y mostrándome en todo momento su amor y apoyo incondicional.

A mi tutor, el profesor Antonio Bula PhD., cuyos conocimientos y dedicación fueron de gran ayuda para realizar este trabajo.

A Alberto Palacio y al profesor Mauricio Carmona PhD., que construyeron con esfuerzo la plataforma experimental de la que obtuve los datos para realizar esta investigación.

A mis amigos, Eduar Pérez, María Díaz, Ovidia Jiménez, Iván Romero, Andrés Rincón, Abraham Sánchez, Rafael Tuirán, Miguelangel Balaguera, Jesús Ortiz, Hugo Álvarez, Freddy Álvarez, Camilo Ramírez, Iván Portnoy, Johnnys Bustillo, Iván Gómez, Francisco Burgos, José Noguera, Blanca Foliaco, Richard Rangel, Elena Romero, Lily Arrieta y Marielena Molinares, por su gran apoyo en cada etapa de mi formación en este posgrado.

Finalmente, a todas las personas que de alguna forma me ayudaron a alcanzar este objetivo, gracias.

Resumen

La energía solar fotovoltaica se encuentra en crecimiento debido a que la eficiencia de los paneles ha aumentado y los costos disminuido en los últimos años. Asimismo, las legislaciones promueven que cada vez sea mayor la capacidad instalada de potencia fotovoltaica. Sin embargo, la naturaleza de la energía solar es intermitente e incontrolable, lo que genera inestabilidad de los sistemas fotovoltaicos que suministran energía a la red. En la literatura se han utilizado desde regresiones multivariables hasta redes neuronales complejas para poder predecir la potencia fotovoltaica. Al desarrollar estos modelos solo se deben incluir las variables independientes que ayudan a explicar el comportamiento de la variable dependiente. Asimismo, los parámetros que se obtienen dependen de la relación entre las variables del modelo. En distintos climas, el efecto de las variables de entrada sobre la variable de salida es diferente, haciendo inválidos los modelos desarrollados en distintos lugares. Al no existir un modelo hecho en el clima del Caribe Colombiano, no se puede conocer con certeza el error que tendría un algoritmo de predicción basado en datos históricos en este clima. Tampoco, qué algoritmo de predicción es el que tiene un mejor desempeño.

En este proyecto se desarrolla una metodología para el ajuste de un modelo de predicción de potencia fotovoltaica utilizando redes neuronales artificiales a partir de datos tomados en la Universidad del Norte, Puerto Colombia. Con el fin de lograr lo anterior, esta investigación se divide en cuatro fases. En la primera se definen las variables importantes, se realiza el tratamiento de los datos y se establece su método de agregación al modelo de predicción de potencia de paneles fotovoltaicos. Para ello se realiza una extensa revisión bibliográfica en la que se estudian los distintos modelos utilizados y el efecto de cada variable ambiental en ellos. En la segunda fase se agrupan los datos con el algoritmo de aprendizaje no supervisado k-means. Como resultado de esta agrupación se obtuvieron dos clústeres, uno de días soleados y otro de días nubosos En la tercera fase se entrenó y validó una red neuronal para cada clúster, y una para el conjunto completo de datos. Esto se hizo ajustando los hiperparámetros de las redes para hallar el menor error de predicción. Para la validación se utilizó una técnica de validación cruzada (10-Fold). Finalmente, la cuarta fase consistió en hacer una comparación entre el modelo propuesto y modelos base de la literatura por medio del nRMSE (distancia media cuadrática mínima normalizada).

Para días soleados, se consiguió un nRMSE de validación de la red neuronal de 5.48%, y para días nubosos el nRMSE fue de 5.24%. El nRMSE del modelo para el conjunto total de datos fue de 5.53%, comprobando que la agrupación de los datos permite tener modelos con menor error. Estos errores se contrastan con el 17.81% del modelo de persistencia. A su vez, se comprobó que para los datos recolectados no fue posible hallar un modelo de regresión lineal múltiple que cumpliera con los supuestos de validación estadística. Los

resultados obtenidos en esta investigación demuestran que la metodología propuesta es de utilidad para disminuir el error de predicción de potencia fotovoltaica.

En este documento se detallan las cuatro fases planteadas en la metodología, así como las conclusiones de la investigación y las recomendaciones a trabajos futuros.

Palabras clave: Predicción, potencia fotovoltaica, energía solar, redes neuronales, k-means, aprendizaje de máquinas.

Tabla de contenido

Lista de figuras

Lista de tablas

Lista de algoritmos

Capítulo 1: Introducción

1.1. Marco conceptual

1.1.1. Tipos de predicción de potencia

Varios enfoques se han tomado para desarrollar modelos de predicción de potencia fotovoltaica. Para entender mejor la estructura de estos distintos enfoques en este documento se plantean las clasificaciones más comunes presentes en la literatura científica del tema. La clasificación de modelos de predicción de potencia se puede hacer según la variable que se predice, según el horizonte de predicción, y según el método utilizado para predecir, como se explica a continuación.

1.1.1.1. Según su variable de predicción

A lo largo de los años en los que se ha buscado predecir valores de potencia de los paneles solares y se han intentado distintos acercamientos para llevar a cabo esta tarea. Inicialmente, los modelos de predicción de potencia se hicieron tomando como variable para predecir la irradiancia, y a partir de distintas expresiones que relacionan la irradiancia con la potencia, llegar a un valor futuro de potencia. Este tipo de predicción se conoce como *predicción indirecta*, puesto hay un proceso de cálculo entre la variable predicha y la potencia de los paneles solares. Cuando el algoritmo de predicción tiene como variable de salida la potencia, al modelo se le denomina de *predicción directa*. A pesar de que las variables meteorológicas se han intentado predecir por mucho más tiempo, y cada vez se cuenta con errores de predicción pequeños, desde hace años, es sabido que la predicción directa entrega modelos con errores menores (Kudo, Takeuchi, Nozaki, Endo, & Jiro, 2009). La ventaja de la predicción indirecta se encuentra en los sistemas recién instalados, donde no se cuenta con datos de producción, pero sí con datos meteorológicos.

1.1.1.2. Según su horizonte de predicción

El horizonte de predicción es esa ventana de tiempo entre el último dato medido y aquel que se desea predecir. Existen las mediciones de *corto*, *mediano* y *largo plazo*, y dependiendo del uso que se le vaya a dar al valor predicho, se recomiendan unos u otros plazos para la predicción. Lastimosamente, no existe a la fecha una regla general que sirva para decir qué tipo de algoritmo entrega mejores resultados en cada uno de los tiempos de predicción (Das et al., 2018). Los tipos de horizonte de predicción son los siguientes.

- Corto plazo: es el pronóstico de la potencia fotovoltaica que se realiza durante una hora, varias horas, un día o hasta siete días. La predicción a corto plazo de la energía fotovoltaica garantiza la programación y el envío de energía eléctrica. Este tipo de predicción es útil para diseñar un sistema

integrado de gestión de energía fotovoltaica. El pronóstico a corto plazo también mejora la seguridad de la operación de la red.

- Mediano plazo: esta predicción es de más de una semana a un mes. Este tipo de pronóstico ayuda a mejorar la planificación de mantenimientos y a predecir la disponibilidad de energía eléctrica en el futuro.
- Largo plazo: los pronósticos de energía fotovoltaica de más de un mes a un año se consideran de largo plazo. La utilidad de tener este tipo de predicciones se encuentra en la mejora de la planificación de generación, transmisión y distribución de la electricidad generada, aparte de la licitación de energía y la operación de seguridad.

Además de los tipos mencionados, los científicos han añadido una cuarta categoría llamada predicción de *muy corto plazo*, y es la predicción que se hace para pocos segundos, un minuto o varios minutos que no sumen más de una hora (Amral, Ozveren, & King, 2007). Este tipo de predicción a muy corto plazo sirve para suavizar el despacho de energía a tiempo real.

1.1.1.3. Según su método de predicción

Las predicciones que utilizan datos históricos de variables ambientales y de energía generada por paneles solares, se pueden dividir según su método o algoritmo de predicción. El método más sencillo es el de *persistencia*. El método de persistencia es utilizado hoy en día como una base para comparar los modelos de predicción hechos con otros métodos más avanzados. Como en el documento de la referencia (da Silva Fonseca et al., 2012). El modelo de persistencia consiste en asumir que el siguiente valor de potencia, será igual al del día anterior a una hora similar. Evidentemente, no todos los días las condiciones climáticas son iguales, y esto lleva a un error, pero el método de persistencia tiene como ventaja su simplicidad matemática.

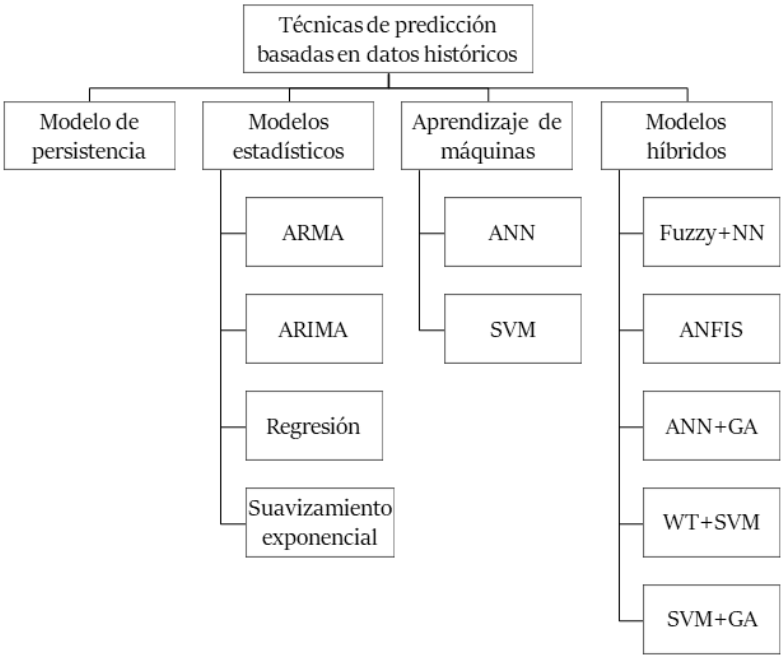

Figura 1 Clasificación de las técnicas de predicción de potencia fotovoltaica basadas en datos históricos. Adaptado de (Das et al., 2018).

Los demás métodos se pueden clasificar como métodos estadísticos, de aprendizaje de máquinas, y métodos híbridos. En la **Figura 1** se muestra un esquema de los distintos métodos de predicción basados en datos históricos, y los algoritmos que hacen parte de cada uno.

ARMA se refiere al modelo autorregresivo de media móvil, y ARIMA al modelo autorregresivo integrado de media móvil y son los métodos más comunes para hacer predicciones de series de tiempo (Abuella & Chowdhury, 2015). En general, estos métodos son adaptables, pueden lidiar con estacionalidad y solo necesitan el último modelo de una serie de tiempo. Pero tienen como desventaja que es poco probable que se desarrollen bien en predicciones a largo plazo, requieren mucho poder computacional y pueden llegar a ser subjetivos, requiriendo alto análisis en la estadística que los fundamenta (Zhai, 2005).

Dentro de los métodos de aprendizaje de máquinas los algoritmos más comunes son redes neuronales artificiales (ANN), máquinas de soporte vectorial (SVM), árboles de decisión (RF), algoritmos genéticos (GA), y algoritmos de agrupamiento (vecinos más cercanos, *k-means*, entre otros). Sin embargo, para desarrollar modelos de regresión que luego sirven para hacer pronósticos de potencia se utilizan con más frecuencia las ANN y SVM. Ahora bien, estos algoritmos se combinan para llegar a modelos con mayor precisión llamados híbridos, como se muestra más adelante en el estado del arte.

1.1.2. Inteligencia artificial

El término *inteligencia artificial* apareció y empezó a ser utilizado en el año 1956 por un grupo de investigadores en una conferencia en el Darmounth College (Moor, 2006). A partir de ese año, se ha convertido uno de los campos de investigación más relevantes. La inteligencia artificial consiste en crear algoritmos o máquinas que sean capaces de imitar comportamientos humanos. Otra definición de la inteligencia artificial es, "La inteligencia artificial (IA) es un sistema informático entrenado para percibir su entorno, tomar decisiones y tomar medidas." (MathWorks, 2020).

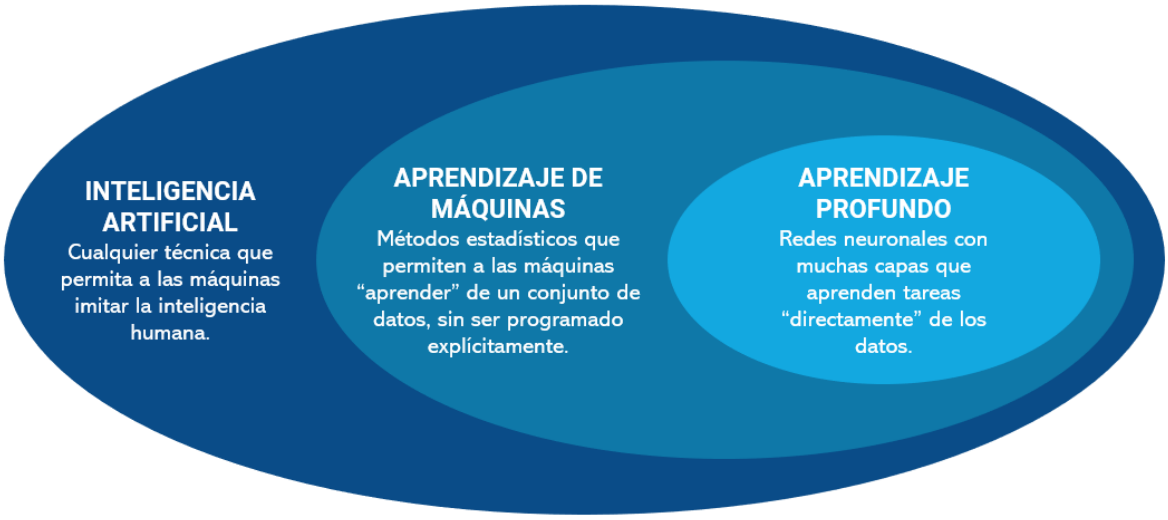

Figura 2 Diferencias entre inteligencia artificial, aprendizaje de máquinas y aprendizaje profundo. Adaptado del curso de Matlab sobre Deep Learning (MathWorks, 2020).

A pesar de que se ha vuelto muy común el uso indiscriminado de los términos inteligencia artificial, *aprendizaje de máquinas* y *aprendizaje profundo*, cada uno tiene un concepto diferente, como se muestra

en la **Figura 2**. En ella vemos que el aprendizaje de máquinas es un tema dentro de la inteligencia artificial. Y el aprendizaje profundo es un tema dentro del aprendizaje de máquinas. El aprendizaje de máquinas se detalla a continuación, en cambio, el aprendizaje profundo tiene aplicaciones cuya complejidad va más allá a la de la predicción de potencia de paneles solares, por lo que no se detalla.

1.1.3. Aprendizaje de máquinas

El aprendizaje de máquinas es un campo de las ciencias de la computación que le da a las computadoras la habilidad de aprender sin ser explícitamente programadas (Samuel, 1959). Samuel enunció en 1959 que una computadora puede ser entrenada con aprendizaje de máquinas de forma que pueda aprender a jugar damas mejor que la persona que escribió el programa (Samuel, 1959). El aprendizaje de máquinas comprende gran variedad de algoritmos que se pueden dividir principalmente en *aprendizaje supervisado* y *aprendizaje no supervisado*. La diferencia entre estos dos se deriva de la existencia de etiquetas en los datos. Los datos están etiquetados cuando hay una distinción entre las variables independientes y la variable de interés o dependiente en un modelo, y además se conoce cuáles variables independientes le corresponden a cada variable dependiente. Cuando se tienen en cuenta estas etiquetas de los datos, se habla de un modelo de aprendizaje supervisado. Cuando no, se habla de un modelo de aprendizaje no supervisado. A continuación, se explican con más detalle estos conceptos.

1.1.3.1. Aprendizaje supervisado

El aprendizaje supervisado se puede definir como una técnica para deducir una función basándose en un conjunto de datos de entrenamiento. Esos datos de entrenamiento se encuentran etiquetados, es decir, se sabe cuáles son las variables de entrada del modelo, y cuál o cuáles son las variables de salida. Los algoritmos de aprendizaje supervisado se pueden utilizar para clasificar datos, o para hacer regresión (por tanto, predicción). En el aprendizaje supervisado se utilizan en general dos grupos de datos. Uno de ellos sirve para *entrenar* el modelo, es decir, hallar los parámetros que minimizan el error. Y el otro sirve para calificar qué tan bien funciona el modelo con distintas medidas de error.

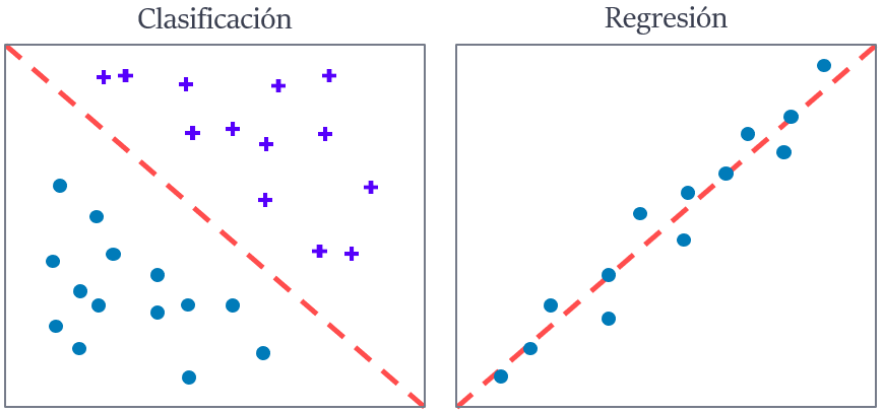

Figura 3 Diferencia gráfica entre clasificación y regresión. Adaptado de (GURU 99, 2020).

Entre los algoritmos de aprendizaje supervisado más comunes se encuentran las ANN, SVM, y GA. Las ANN (o simplemente NN) son técnicas populares de aprendizaje automático que simulan el mecanismo de aprendizaje en organismos biológicos (Aggarwal, 2018). Se componen de neuronas conectadas entre sí en

distintos niveles o capas. Las neuronas reciben información de la capa anterior y las pasan a la siguiente capa, en esto se parecen a las ANN. Una neurona se puede describir como un modelo de regresión lineal múltiple, debido a que internamente hace una suma ponderada de las variables de entrada, al multiplicarlas por unos pesos. Adicional a esto, suma un término al que se le llama *sesgo*. A la salida de la suma, se encuentra una *función de transferencia* que se encarga de añadir la no linealidad de la neurona. Gráficamente se puede ver en la **Figura 4**.

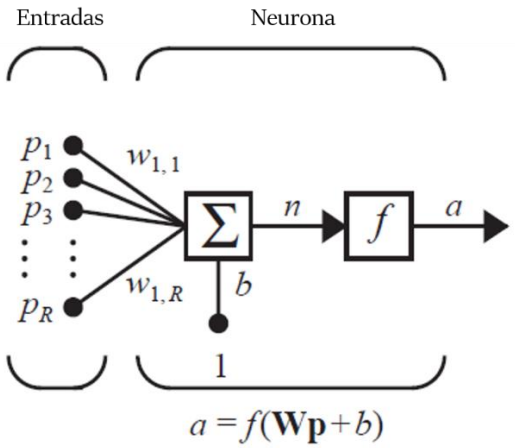

Figura 4 Neuronas de múltiples entradas. Adaptado de (Hagan & Demuth, 2014).

Las neuronas se pueden ubicar en capas, que se conectan entre sí para formar una red. En la **Figura 5** se muestra una red neuronal de tres capas. Las salidas de una capa son las entradas de la siguiente.

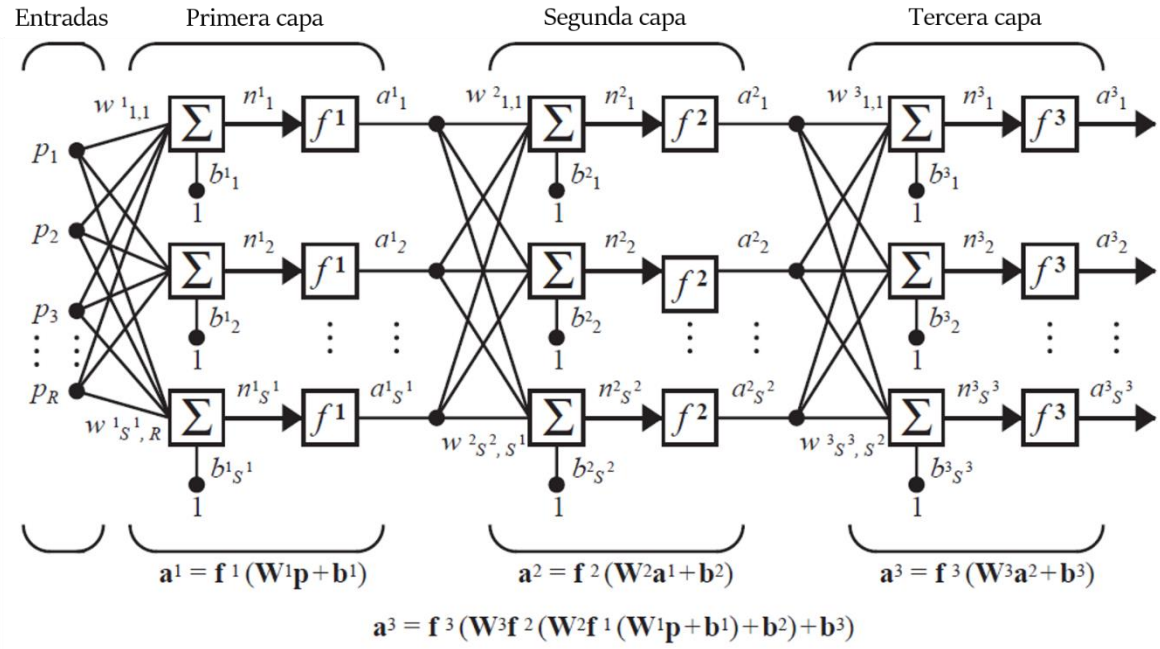

Figura 5 Red neuronal de tres capas. Adaptado de (Hagan & Demuth, 2014).

Los pesos por los cuales se multiplican los valores de entrada son valores por ajustar. Los valores por ajustar en una red neuronal reciben el nombre de *parámetros*. El algoritmo de optimización más popular para hallar el mejor valor de esos pesos es conocido como *propagación hacia atrás* (BP). Las funciones de transferencia, el número de capas de la red, número de neuronas por capa, entre otras configuraciones de la red, son llamados *hiperparámetros*, y sus valores se eligen al iniciar el entrenamiento de la red. Las funciones de transferencia más conocidas son la sigmoidal y la tangente hiperbólica. Las funciones sigmoidales restringen la salida de la neurona entre 0 y 1, a diferencia de una función lineal cuyo rango es desde infinito negativo hasta infinito positivo. La función tangente hiperbólica restringe la salida de la neurona entre -1 y 1. Estas restricciones añaden cierta estabilidad a la red neuronal. La función sigmoidal tiene la siguiente forma,

$$A = \frac{1}{1 + e^{-x}} \qquad \textbf{Ec. 1}$$

Mientras que la función tangente hiperbólica es de la forma,

$$A = \frac{e^{n} - e^{-n}}{e^{n} - e^{-n}} \qquad \textbf{Ec. 2}$$

Entre otras funciones de transferencia se encuentran la lineal, la lineal saturada y la de pulso. La forma de estas se verá más adelante en el **Capítulo 4** de este documento.

1.1.3.2. Aprendizaje no supervisado

Por otro lado, el aprendizaje no supervisado no tiene etiquetas en los datos. Por ende, todos los datos son tratados igual. Lo que se busca en estos modelos es encontrar grupos o clasificaciones. Uno de los más populares de estos algoritmos es el de *k-means*. En donde k es el número de *clústeres* o grupos en los que se desea dividir la información. Dentro de un clúster los valores son parecidos entre sí y diferentes a los valores de otros clústeres. En la **Figura 6** se muestra la función de un algoritmo de agrupación para datos bidimensionales.

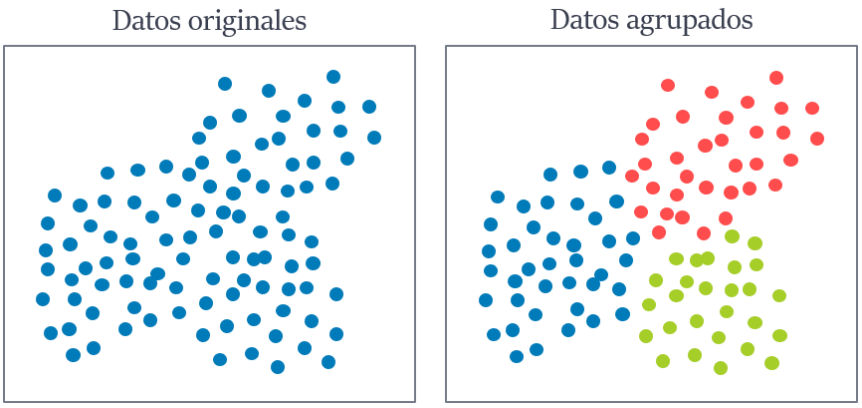

Figura 6 Ejemplo de un algoritmo de agrupación. Adaptado de (PyRP, 2020).

El algoritmo de k-means funciona al ubicar k centros de forma aleatoria, luego mide la distancia euclidiana de cada punto a ese centro, y se ubican los puntos en el grupo del centro más cercano. Una vez hecho esto, calcula un nuevo centroide y repiten los pasos anteriores hasta que los centroides no tengan cambios.

1.1.4. Error de predicción

El error de predicción es la forma en la que se mide el rendimiento de un modelo. También sirve para poder comparar un modelo de predicción con los demás. Existen varias formas de medirlo, pero en la siguiente tabla se muestran los más comunes presentes en la literatura por sus siglas en inglés.

Tabla 1 Fórmulas para los distintos tipos de error de predicción.

Siglas	Nombre	Fórmula	
MSE	Error cuadrático medio.	$\frac{1}{N}\sum_{i=1}^{N}\left(y_{pred}-y_{med}\right)^2$	**Ec. 3**
RMSE	Raíz del error cuadrático medio o distancia media cuadrática mínima.	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{pred}-y_{med}\right)^2}$	**Ec. 4**
nRMSE	Distancia media cuadrática mínima normalizada.	$\left(\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_{pred}-y_{med}\right)^2}\right)\cdot 100\%/y_{med_{máx}}$	**Ec. 5**
MAE	Error absoluto medio.	$\frac{1}{N}\sum_{i=1}^{N}\left\|y_{pred}-y_{med}\right\|$	**Ec. 6**
MAPE	El error porcentual absoluto medio.	$\frac{1}{N}\sum_{i=1}^{N}\frac{\left\|y_{med}-y_{pred}\right\|}{\left\|y_{med}\right\|}\cdot 100\%$	**Ec. 7**
MRE	Error medio relativo.	$\frac{1}{N}\sum_{i=1}^{N}\frac{y_{pred}-y_{med}}{y_{total}}\cdot 100\%$	**Ec. 8**
MBE	Sesgo promedio.	$\frac{1}{N}\sum_{i=1}^{N}\left(y_{pred}-y_{med}\right)$	**Ec. 9**

Donde N es el número de observaciones, y_{pred} es la potencia predicha por el modelo, y_{med} es la potencia medida, $y_{med_{máx}}$ es el valor máximo medido de potencia. Finalmente, y_{total} es la capacidad instalada del sistema fotovoltaico. Cabe resaltar que el RMSE es frecuentemente utilizado en la literatura académica a pesar de no ser un porcentaje. En los modelos de potencia fotovoltaica el RMSE tiene las unidades de la potencia.

1.2. Estado del arte

La energía solar es una de las fuentes renovables más importantes y estudiadas en la actualidad. Hoy en día, son muchos los países que promueven la utilización de ésta tanto a escala industrial como doméstica. A esta lista de países se sumó Colombia en el año 2014, con la ley 1715 de ese año, que estimula las fuentes no convencionales de energía (FNCE) tanto en el Sistema Interconectado Nacional como en las zonas no interconectadas (ZNI) de Colombia (Lubo, 2019). Con esta ley se fortalecieron los esfuerzos que se habían

estancado en años anteriores para incentivar la apropiación de la energía fotovoltaica al conceder beneficios a personas naturales o jurídicas que fomenten la investigación, desarrollo e inversión en el ámbito de la producción y utilización de energía a partir de FNCE.

Desde que se fabricó el primer panel solar con eficiencia del 1% en 1883, a cargo de Charles Fritts (Fritts, 1883), los investigadores se han centrado en cambiar los materiales de los paneles para aumentar su eficiencia y por ende obtener mayor producción de energía. En el año 1960, el máximo de eficiencia fue alcanzado por Hoffman Electric, siendo un 14% (Matasci, 2018). En 1992, University of South Florida fabricó celdas de pared delgada excediendo el 15% de eficiencia por primera vez con eficiencia de 15.89% (Cleveland & Morris, 2014). Luego, en el 2012 Solar Frontier alcanzó 17.8% de eficiencia y tres años después First Solar y Sun Power superaron ese límite consiguiendo 18.2% y 22.8% de eficiencia, respectivamente. En el 2017 se desarrolló una celda capaz de entregar el 44.5% de eficiencia. Hasta la fecha, la eficiencia más alta validada por el National Renewable Energy Laboratory (NREL) es de 47.1% (Geisz et al., 2020). Cabe aclarar que estas eficiencias elevadas se deben al uso de materiales diferentes a silicio que suelen ser más costosos y no permiten que la tecnología sea rentable. En el mercado actual, se cuenta con paneles de silicio que tienen eficiencias de entre 15 y 20%, y alcanzan en promedio los 20 años de vida útil (Vikram, 2020).

De igual forma, ahora existen muchos modelos algebraicos que proporcionan información sobre la relación entre variables ambientales y la potencia generada por los paneles, la mayoría de éstos incluye el efecto de la irradiancia, temperatura de la celda, ensuciamiento de los paneles o nubosidad. Sin embargo, la naturaleza de esta generación eléctrica es intermitente e incontrolable, lo que hace que sea difícil depender de ella. Esta desventaja de los paneles solares se ha intentado eliminar con el uso de modelos de predicción que permitan brindar mayor estabilidad a la red eléctrica a la que se conectan.

La revisión bibliográfica presentada en este trabajo se hizo al combinar los siguientes términos de búsqueda y términos afines a éstos: “photovoltaic*”, “power*”, y “forecasting*”. Estos términos se buscaron en el título de documentos contenidos en la base de datos Web of Science de Clarivate Analytics. Obteniendo el conjunto de resultados en donde coinciden los temas de los términos utilizados. La búsqueda de documentos se realizó el 25 de mayo del 2020, limitando la búsqueda a documentos publicados a partir del año 2000. Datos de estadística descriptiva de los resultados se encuentran en la **Tabla 2**. En total, se obtuvieron 855 documentos que incluyen artículos científicos, artículos de revisión, resultado de conferencias, artículos de acceso temprano, entre otros.

Tabla 2 Análisis descriptivo de la colección de documentos.

Descripción	Resultados
Documentos	855
Periodo	2000 - 2020
Porcentaje de crecimiento anual	20.97%
Promedio de citaciones por documento	10.56
Autores	2352
Apariciones de los autores	3342
Autores de documentos de única autoría	18
Autores de documentos de múltiple autoría	2334

Documentos por autor	0.364
Autores por documento	2.75
Coautores por documento	3.91
Índice de colaboración	2.81

El número de documentos por autor es igual al número de documentos total entre el número de autores. En cambio, el número de autores por documento es el inverso de los documentos por autor. Por otro lado, el índice de colaboración se define como el número promedio de autores por artículos de múltiple autoría (Elango & Rajendran, 2012). En la **Figura 7** se muestra el incremento del número de publicaciones desde el año 2000 hasta el presente en la temática de modelos de predicción de potencia de paneles solares fotovoltaicos. La línea muestra el número total de publicaciones hasta este año, o el acumulado de publicaciones, y el histograma muestra el total de citaciones para cada año. Desde el año 2011 se ha notado un incremento a gran velocidad en el número de publicaciones en el que, al excluir el año 2020 por no haber terminado, es notorio que no hay una inclinación hacia meseta en los últimos años. Esto da indicios de que el tema aún se encuentra en estado de desarrollo y la investigación en él sigue siendo pertinente.

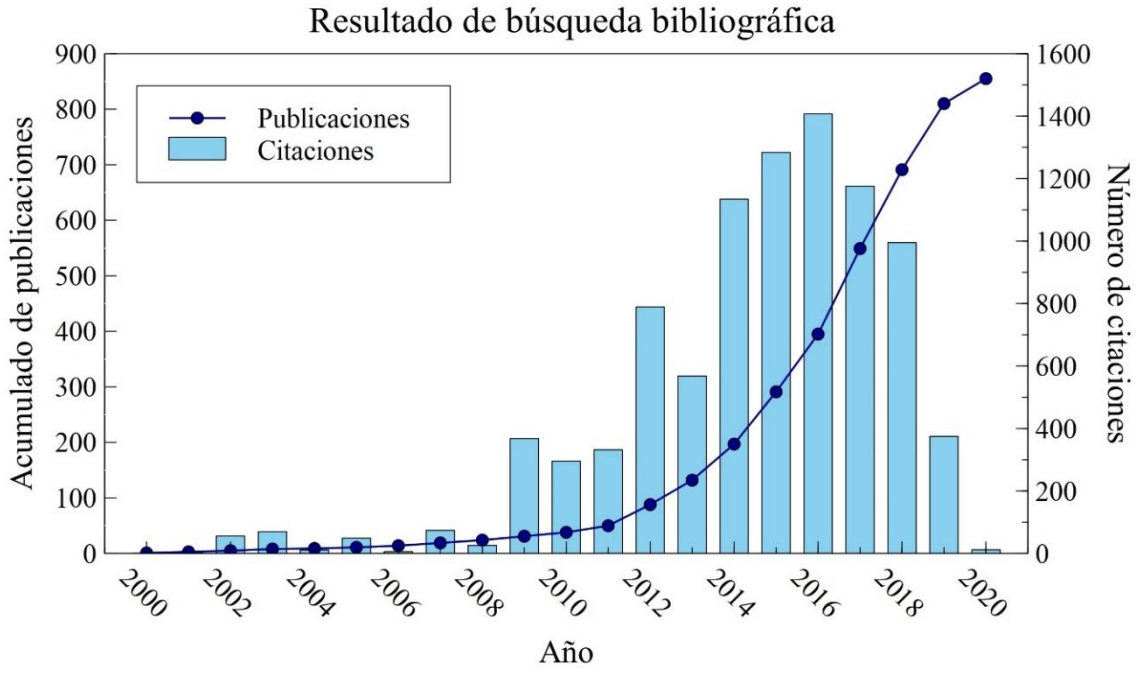

Figura 7 Acumulado del número de publicaciones y número de citaciones de modelos de predicción de paneles fotovoltaicos desde el 2000.

Estudios de bibliometría han sido incluidos en este trabajo, utilizando como herramienta de analítica de información el paquete Bibliometrix de RStudio® y su plataforma web BiblioShiny (Aria & Cuccurullo, 2017). De esta herramienta se puede obtener información que a simple vista no serían tan evidente sobre la colección de documentos que se ha encontrado en el tema a investigar. Es por eso, que en los siguientes párrafos se mostrará una breve discusión sobre los distintos métodos que sirven para conocer a profundidad la forma en la que se encuentran distribuidos los temas, la dirección en la que los investigadores han centrado sus trabajos y la estructura conceptual e intelectual de la colección de

documentos. Al final, se concluye con las brechas tecnológicas encontradas en el estado del arte y cómo este documento intenta cubrir parte de ellas.

El análisis de las contribuciones de cada artículo se puede llevar a una mayor profundidad con la espectroscopia de las referencias, que se aprecia en la **Figura 8**. Esto es, la frecuencia con que se citan las referencias en las publicaciones de un campo de investigación específico (Rhaiem & Bornmann, 2018), no en todos los campos de investigación. Esta gráfica ayuda a conocer en qué año se dan los avances más significativos en el campo que se estudia. Con este método, se pueden determinar las raíces históricas de los campos de investigación y cuantificar su impacto en la investigación actual (Marx, Bornmann, Barth, & Leydesdorff, 2014). Por ejemplo, para esta investigación bibliométrica, entre los años 2013 y 2015 se puede decir que hubo un avance significativo en el área de interés de este trabajo.

Figura 8 Espectroscopia de las referencias por año. En celeste las desviaciones de la media de cinco años.

En el año 2013 hubo 1953 citaciones en las publicaciones de predicción de potencia fotovoltaica, en el 2014 y en el 2015 hubo 1913 y 1971, respectivamente. Gran número de las citaciones del 2013 corresponden al artículo de revisión de Inman, Pedro y Coimbra, en el que se sugiere la utilización de modelos híbridos por su reducción en el error de predicción (Inman, Pedro, & Coimbra, 2013). También influyó al avance el artículo de Diagne et al. en el que se hace una revisión literaria que resalta de igual manera la ventaja del uso de modelos híbridos. Por otro lado, destaca que el uso de métodos de inteligencia artificial presenta mejores resultados de predicción que los métodos de regresión tradicionales (Diagne, David, Lauret, Boland, & Schmutz, 2013). De igual forma, en ese año se presentaron avances por parte de los autores Bouzerdoum, Mellit y Pavan, quienes desarrollaron un modelo híbrido entre el método autorregresivo integrado de promedio móvil estacional (SARIMA) y SVM (Bouzerdoum, Mellit, & Massi Pavan, 2013). Con este modelo, demostraron que el rendimiento del híbrido entre los dos métodos para predicciones de una hora arroja mejores predicciones que cada método independiente. El nRMSE para este modelo fue de

aproximadamente 9.4% destacando como ventaja de este modelo que no se necesitan predicciones de variables ambientales para predecir potencia. Por otro lado, los autores Pelland, Galanis y Kallos desarrollaron un modelo de predicción de potencia de forma indirecta, al predecir factores ambientales (Pelland, Galanis, & Kallos, 2013). El nRMSE de este modelo de predicción de irradiancia global estuvo entre 6.4% y 9.2%.

Durante el 2014 la tendencia de los modelos híbridos continuó con el trabajo de Yang et al. en el que presenta un modelo que predice potencia hasta un día adelante, de forma horaria (Yang, Huang, Huang, & Pai, 2014). El método híbrido se descompone en etapas de clasificación, entrenamiento y predicción. Para la clasificación, los autores utilizaron mapa de autoorganización (SOM) y las redes de cuantificación del vector de aprendizaje (LVQ) se utilizan para clasificar los datos históricos recopilados de la producción de energía fotovoltaica. En el entrenamiento emplearon la regresión de vectores de soporte (SVR) y en la predicción, la inferencia difusa. Obtuvieron un modelo con RMSE de 350.2 W y RME de 3.29%. Li et al. desarrollaron un modelo que, a diferencia del ARIMA, sí utiliza datos meteorológicos para hacer una predicción con más información y exactitud (Li, Su, & Shu, 2014), dando como resultado un modelo con RMSE de 125.84 W. Modelos de predicción de corto plazo con redes neuronales también fueron publicados ese año, como es el caso de los autores De Giorgi, Congedo y Malvoni, (De Giorgi, Congedo, & Malvoni, 2014). Adicional al modelo de predicción, los autores pudieron determinar por medio de un análisis de sensibilidad que la irradiancia es la que tiene mayor efecto sobre la potencia generada. También fue presentado el modelo de Mellit et al., que predice potencia utilizando valores futuros de irradiancia y temperatura de los paneles con distintos modelos de redes neuronales para cada tipo de día. Obtuvieron un error MAPE mínimo del 1.92%.

Finalmente, el 2015 se publicó un trabajo incluyendo por primera vez el índice de aerosol, que indica la cantidad de partículas en la atmósfera. Estas partículas absorben la energía solar disponible, por lo que el índice de aerosol afecta la generación de energía fotovoltaica, específicamente con un índice de correlación de -0.32 para el caso estudiado en el documento (J. Liu, Fang, Zhang, & Yang, 2015). Liu et al. desarrollaron modelos de predicción con redes neuronales con y sin el índice de aerosol, llegando a que se reduce el error al incluir esta variable.

Para conocer los subtemas en los que se encuentra dividido el campo de investigación, estudiamos el mapa de estructura conceptual que se encuentra en la **Figura 9**. La estructura conceptual trata de explicar los principales temas y tendencias en el mundo científico, es decir, de qué habla la ciencia. En la colección se pueden ver 3 clústeres (o temas) que se han definido por las palabras clave dadas por los autores. Para entender la temática asociada a cada clúster, es necesario estudiar los documentos que se encuentran en ellos. Los tres clústeres o temas tienen en común el tema global, que son algoritmos de predicción de potencia fotovoltaica, ya sea de forma directa o indirecta. Los cambios en los clústeres tienen que ver con los métodos que se utilizan para llegar a estos modelos de predicción. En el primer clúster se encuentran documentos de predicción de potencia utilizando algoritmos de redes neuronales artificiales en su mayoría y de máquinas de soporte vectorial, que son los algoritmos más comunes. La temática del segundo es muy parecida a la del primer clúster, sin embargo, este no incluye los métodos de SVR y SVM en la predicción de potencia. El clúster 3 en verde, tiene en cambio como tema central la predicción de potencia por medio de modelos de control predictivo (MPC). Esto se ve en la investigación de Mosa et al. en donde implementan

este tipo de modelo como ayuda del algoritmo de seguimiento del punto de máxima potencia (MPPT) (Mosa, Shadmand, Balog, & Rub, 2017).

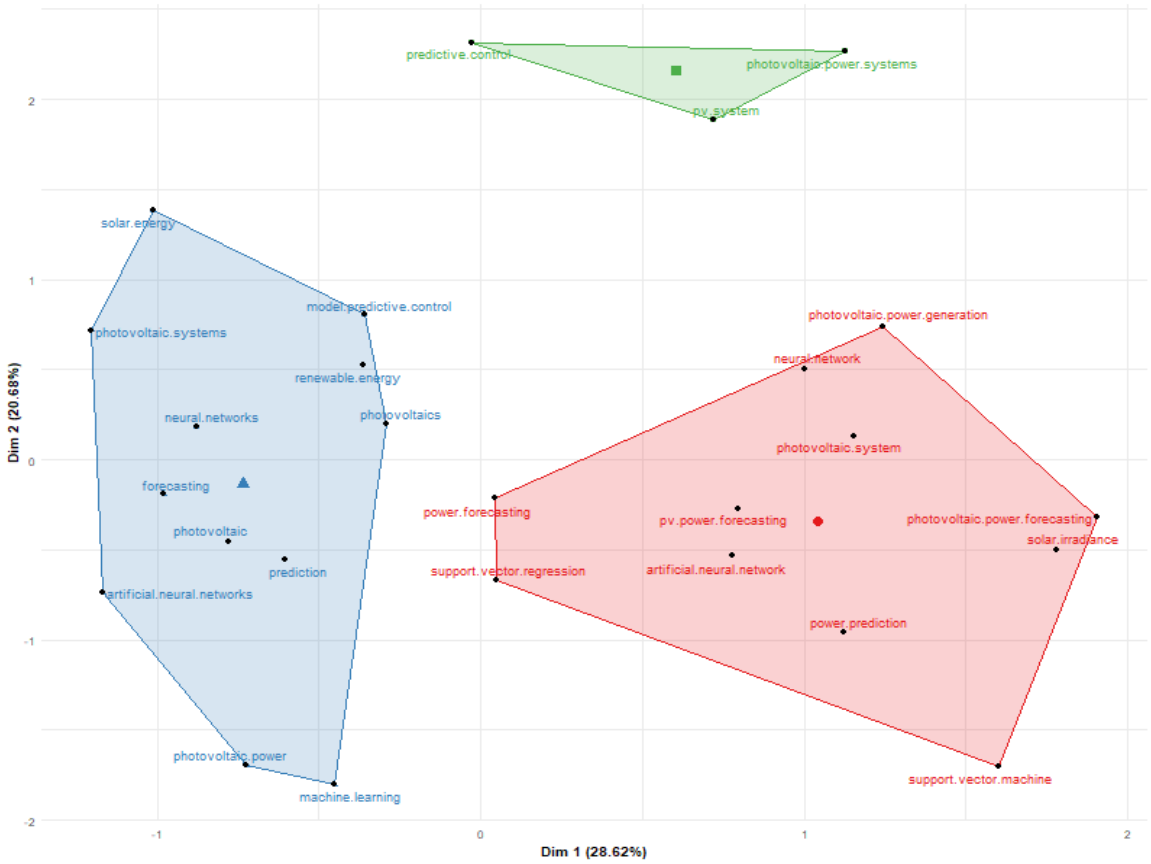

Figura 9 Mapa de estructura conceptual de la colección bibliográfica.

Una ayuda visual para entender la estructura conceptual del tema de investigación es el dendograma de los temas. En la **Figura 10** se encuentra la distribución de las palabras clave, agrupadas por tema y con el mismo color de los clústeres en la **Figura 9**. La altura de las líneas que las conectan indica la proximidad de estas palabras, mientras más baja es la línea que conecta dos palabras, más cercanas son estas entre sí. Si dos palabras son utilizadas juntas en gran cantidad de documentos, entonces serán próximas. En cambio, si en pocos documentos se utilizan las dos palabras juntas, se encontrarán distantes en este diagrama. Se puede ver como algunas palabras aparecen en varios clústeres, con pequeñas variaciones en la escritura, pues cada autor las escribe de una forma diferente. A pesar de ello, es fácil ver la diferencia entre los temas de los clústeres con este diagrama.

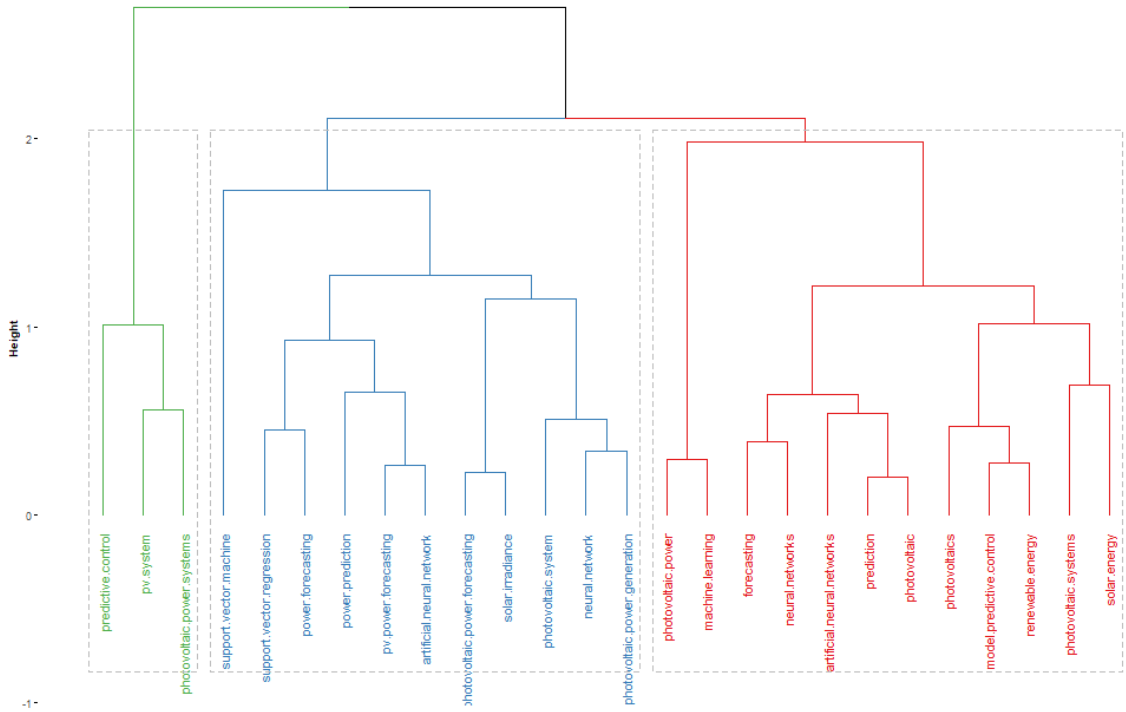

Figura 10 Dendograma del análisis factorial para la colección bibliográfica.

Por otro lado, en la evolución temática de la colección de datos que se encuentra en la **Figura 11**, se muestra cómo las palabras clave de los autores han ido cambiando en el tiempo. Llama la atención como el término "prediction" es reemplazado por "forecasting". De igual forma, como hoy en día "neural networks" es de los términos más comunes al referirse a la predicción de potencia de paneles solares. Similarmente, se ve como el término "predictive control", presente en el clúster 3, apareció en el 2014 y sigue presente en los documentos actuales. Por último,el término "support vector regression" se transforma en "forecasting" y "neural networks".

Estos términos se ven con gran frecuencia en las investigaciones recientes, donde sobresale el uso de las redes neuronales por su repetición en los distintos documentos. En (Colak, Yesilbudak, & Bayindir, 2020) se compararon varios algoritmos de optimización de redes neuronales y varias funciones de activación llegando a que la función de activación sigmoide se desempeña mejor en la predicción de potencia. Redes neuronales de corta y larga memoria (LSTM) fueron utilizadas para predecir potencia de paneles solares y de energía eólica por (Han et al., 2019) a mediano y largo plazo. Los errores obtenidos fueron menores a los del modelo de persistencia y un modelo hecho con SVM. Un método novedoso fue presentado por (Z. Liu, Li, Tseng, & Lim, 2020) en el que se utiliza un modelo inspirado en un enjambre de gallinas buscando alimento para optimizar un modelo basado en redes neuronales para predecir potencia a corto plazo. Estos autores clasificaron los datos en días soleados, lluviosos o nubosos, y obtuvieron un nRMSE promedio de 5.54% para sus modelos.

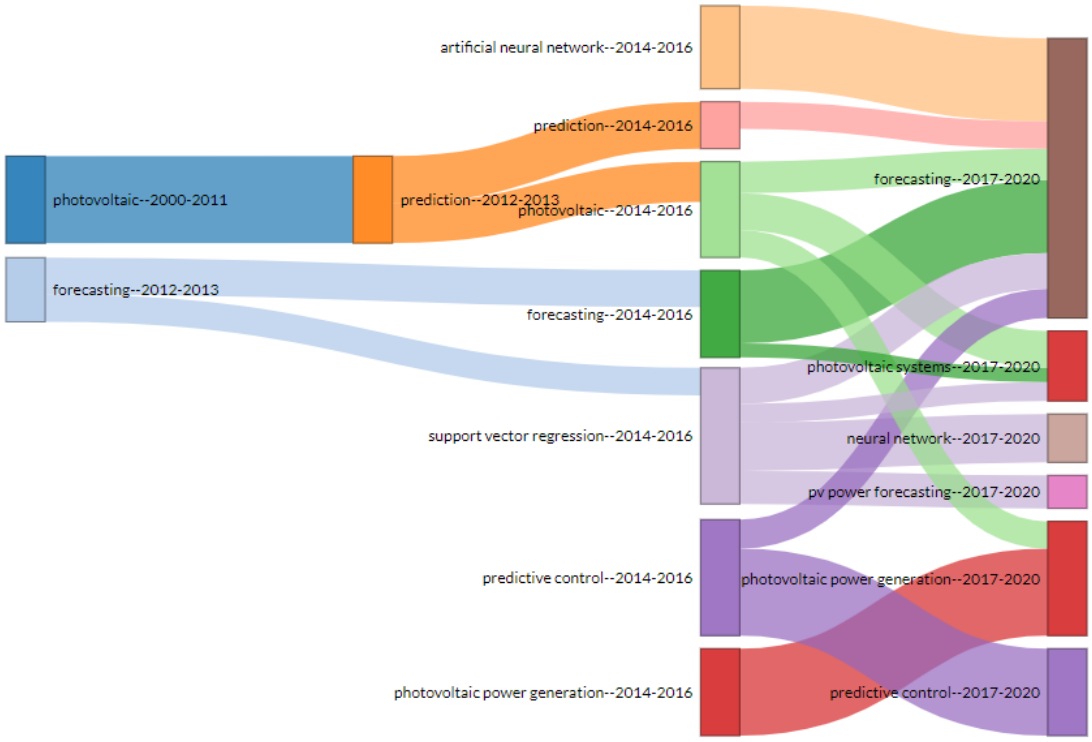

Figura 11 Evolución temática dividida en cuatro periodos.

A lo largo de este estudio se han descrito distintos trabajos con alto impacto en la comunidad científica. En varios de estos documentos se encuentran barreras por mejorar y brechas de conocimiento por cubrir. Muchas de éstas ya han sido tratadas, sin embargo, otras requieren mayor desarrollo de las tecnologías para ser cubiertas por completo. A continuación, se describen las principales brechas encontradas en este trabajo:

- En la mayoría de los artículos se busca hacer predicciones a corto plazo debido a que los errores de predicción siguen siendo muy grandes para los pronósticos a largo plazo (Das et al., 2018).
- A pesar de ser una herramienta muy utilizada en la predicción de energía fotovoltaica por su capacidad de predicción en casos donde la aleatoriedad de los datos es alta, las ANN necesitan muchos datos para poder llegar a resultados buenos. Además, son pocos los casos en los que se hacen análisis para conocer el efecto de cada variable en el modelo.
- En la costa colombiana no hay trabajos en los que se haya hecho predicción de potencia directa. La mayoría de los trabajos proviene de China (452 artículos), seguido por Estados Unidos (203 artículos), Japón (125 artículos) e Italia (103 artículos). Todos son países de alta latitud, cuyos climas son diferentes al de Colombia.
- En general, los documentos utilizan el método de agrupación de vecinos cercanos, que no forma grupos, sino que les asigna valores nuevos a las clases que ya se han definido. El algoritmo de k-means en cambio busca crear un número óptimo de clústeres acorde con la información con la que se cuente.

1.3. Planteamiento del problema y justificación

La energía solar fotovoltaica se encuentra en un gran crecimiento debido a que hace algunos años los precios del kilovatio por hora han disminuido, haciendo los paneles solares más atractivos desde el punto de vista económico. La mayoría de los paneles comerciales están hechos de silicio y sus costos en el tiempo se han disminuido por la mejora en los procesos de manufactura de sus fabricantes. La ley de Swanson establece que el precio de los paneles disminuye aproximadamente un 20% cada vez que se doblan las ventas de la producción mundial de paneles solares, por lo que se espera que el precio siga en descenso. En la **Figura 12** se muestra la evolución del precio del Vatio producido por paneles solares en USD por semestres. H1 y H2 son el primer y segundo semestre del año, respectivamente.

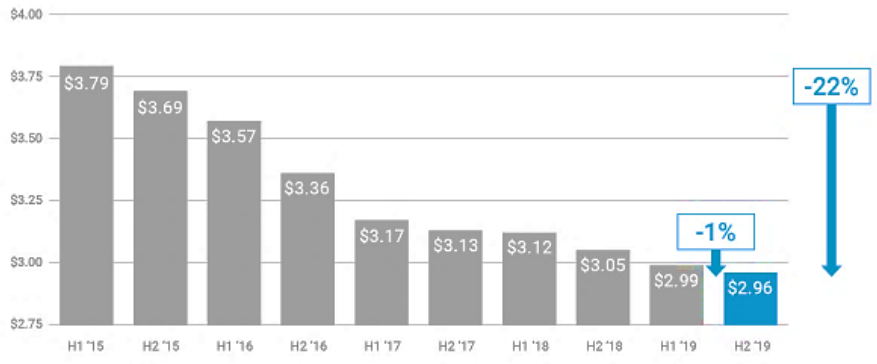

Figura 12 Evolución del precio del Vatio producido por paneles solares en USD por semestres. La gráfica va desde el primer semestre del 2015 hasta el segundo semestre del 2019. Tomado de (energysage, 2019).

La energía fotovoltaica se encuentra ligada a la variabilidad del clima. Por años se han utilizado baterías para mitigar este efecto. Sin embargo, las baterías son componentes con alto valor que hacen que el costo de la energía producida aumente su valor. Con el aumento global del número de plantas fotovoltaicas y las distintas acciones gubernamentales que promueven la instalación de más fuentes de energía fotovoltaicas, nuevas alternativas son necesarias. Por ello, en los últimos años los científicos han trabajado en desarrollar modelos de predicción de potencia fotovoltaica. Tales predicciones son necesarias para programar funciones, así como para despacho económico, control predictivo de frecuencia, análisis de seguridad, restauración de sistemas y comercio de energía (Amral et al., 2007).

Como se enumeró en el estado del arte, gran cantidad de artículos han sido publicados intentando predecir potencia de paneles solares para mitigar este problema, sin embargo, existen brechas en la literatura por cumplir, como la ausencia de este tipo de estudios en Colombia. Uno de los métodos para elegir las variables que se utilizan en un modelo es el análisis de correlaciones. En este se seleccionan como variables influyentes aquellas cuyos índices de correlación con la variable dependiente sea alto. Para el caso presentado en (Das et al., 2018), la correlación entre la temperatura ambiente y a potencia es de 0.38, en cambio, en el caso de estudio de esta investigación, esta correlación es de 0.73. Esta gran diferencia indica que sería un error tomar un modelo desarrollado para un clima diferente, puesto que la relación entre variables meteorológicas y potencia del sistema fotovoltaico difiere en gran manera. Variables que son consideradas en otros modelos pueden no tener efecto en la potencia para el clima en Puerto Colombia. Lo

contrario también puede ocurrir, que variables que en otros climas no tengan efecto en la potencia, en el clima caribeño sí lo tengan, y por error no se incluyan en el modelo. Es por esto por lo que en esta investigación se propone un modelo híbrido de predicción de potencia con datos colectados en la Costa Caribe Colombiana, utilizando ANN y agrupación con k-means.

1.4. Objetivos

1.4.1. Objetivo general

- Desarrollar una metodología de ajuste de modelos de predicción de potencia de un sistema solar fotovoltaico por medio de agrupación con k-means y redes neuronales artificiales para reducir el error de predicción.

1.4.2. Objetivos específicos

- Definir variables importantes y su método de agregación para determinar la estructura del modelo de predicción de potencia de paneles fotovoltaicos.
- Establecer clústeres de datos que permitan obtener modelos con mejor capacidad de predicción.
- Determinar la estructura de los modelos de regresión con redes neuronales para la predicción de potencia de cada uno de los clústeres identificados.
- Analizar la mejora en el rendimiento de los modelos obtenidos al comparar con modelos base encontrados en la literatura.

1.5. Metodología

Con el fin de lograr el objetivo general de este proyecto, se han designado cuatro objetivos específicos. Cada objetivo específico se relaciona con distintas actividades que, al ser culminadas, le dan cumplimiento a cada objetivo. Por simplicidad, al momento en el que se realizan las actividades de cada objetivo específico, se le llamarán fases. A continuación, se detallan las actividades que se deben realizar en cada una de las fases del proyecto y su orden de ejecución.

Fase 1: Definición de las variables y tratamiento de éstas.

Al iniciar el proyecto se llevó a cabo una revisión bibliográfica exhaustiva para validar la importancia del proyecto y encontrar las brechas tecnológicas que se planean cubrir con la realización de éste. Para ello se utilizó el componente Biblioshiny, perteneciente a la herramienta Bibliometrix del software libre RStudio®. Los textos por analizar con esta herramienta se obtuvieron de la base de datos Web Of Science de Clarivate Analytics. Asimismo, como resultado de la revisión bibliográfica se encuentran las variables que han sido utilizadas en modelos similares y se pueden elegir las que se utilizarán en el modelo de predicción. Para el modelo de predicción se utilizaron datos provenientes de una estación meteorológica y de una plataforma experimental con un sistema fotovoltaico, ambos equipos ubicados en la Universidad del Norte. Las estaciones meteorológicas entregan gran cantidad de variables que no están relacionadas necesariamente con el fenómeno fotovoltaico por el cual se rigen los paneles solares. Por ello se hizo de gran importancia

obtener del análisis de la bibliografía las variables que son estudiadas con frecuencia y han demostrado tener efecto en la producción de energía fotovoltaica.

Por otro lado, antes de poder entrenar un algoritmo de inteligencia artificial casi siempre es necesario hacer una etapa de procesamiento de los datos para que sean de utilidad en el modelo (Loy, 2019). Para hacer modelos de regresión con redes neuronales, se necesita que todos los datos que se utilizan sean numéricos. Para ello se debe tratar con los NA (datos que no existen o *not a number*) que se encuentren en los sets de datos. Esto se puede hacer eliminando las filas de datos en las que se hallen valores faltantes, o bien, imputando datos. Si bien existen soluciones rápidas, como la sustitución de la media, que pueden estar bien en algunos casos, tales enfoques simples generalmente introducen sesgo en los datos. Por ejemplo, la aplicación de la sustitución de la media deja la media sin cambios (lo cual es deseable) pero disminuye la varianza, que puede ser indeseable. Cada una de las opciones anteriores tiene sus ventajas y desventajas, y dependiendo del número y distribución de los NA se debe evaluar cuál de estas elegir. En este documento se hizo sustitución de los datos faltantes por medio de interpolación lineal.

Adicional a esto, se evaluaron los índices de correlación de las variables independientes con la potencia fotovoltaica para elegir cuáles de estas debían incluirse en el modelo de predicción. Es decir, se utiliza el análisis de correlaciones para determinar las variables que se encuentran las variables con alta correlación con la potencia de los paneles y poder descartar aquellas que tienen una baja correlación.

El software en el que se programó el algoritmo para hallar los parámetros de los modelos es RStudio®. Este software es especializado en la analítica de datos y cuenta con gran cantidad de funciones que facilitan la realización de diversos cálculos. Sin embargo, este software realiza todas sus operaciones en la RAM del dispositivo, lo que exige que el número de datos con los que se va a trabajar deba elegirse adecuadamente. Esto se tuvo en cuenta en la elaboración del método de agregación de variables del modelo.

Fase 2: Clasificación de los datos.

Ahora bien, la segunda fase consiste en clasificar los datos de forma tal que los datos pertenecientes a cada clasificación o grupo se parezcan entre sí, pero sean poco similares a los datos en otro grupo. Para ello se utilizó el método de agrupación k-means, el cual es un método de aprendizaje de máquinas no supervisado que consiste en dividir los datos en "k" grupos. Antes de ejecutar algoritmos de clasificación, también se hizo necesario normalizar los datos de forma que tuvieran media cero y desviación estándar uno. Uno de los parámetros de entrada de este método es el valor de k, o número de grupos, y existen distintos métodos que ayudan a elegir el valor óptimo de k para garantizar que efectivamente las observaciones en el mismo grupo sean similares y las observaciones en diferentes grupos sean diferentes, sin llegar a un número innecesariamente grande de clústeres. Entre estos métodos están el *método del codo* (elbow method) y *método de silhouette* (silhouette method). Ambos fueron utilizados en este trabajo para determinar el k óptimo. Luego de haber hallado el número de clústeres óptimo y los datos fueron agrupados, se crearon distintos conjuntos de datos para cada clúster y se procedió a estudiarse el contenido dentro de cada clúster. Esto se hizo obteniendo el valor promedio de cada variable en los distintos clústeres y graficando la irradiancia y la potencia promedio en cada clúster para conocer sus valores mínimo, máximo y su nueva tendencia. Similarmente, se estudiaron las correlaciones entre las variables de entrada y la variable de salida para verificar que las variables a incluir en cada modelo fueran las adecuadas.

Fase 3: Entrenamiento y validación de las redes neuronales.

Para entrenar las redes neuronales se utiliza el algoritmo de retropropagación resiliente con retroceso de pesos. De igual forma, se ajustan los hiperparámetros de forma que se obtenga una red neuronal con bajo error de predicción, esto es, elegir distinto números de neuronas por capa, para comparar las distintas arquitecturas de red neuronal y elegir aquella que presente menor error de predicción. La métrica utilizada para la elección del conjunto de hiperparámetros óptimo fue el RMSE. Dentro de los requerimientos para obtener un buen modelo de predicción con redes neuronales está que no se presente sobreajuste de las redes. El sobreajuste se da cuando una red encaja tanto en los datos de entrenamiento, que se ajusta a su ruido, por lo que cuando se le presenten nuevos datos, probablemente con un nivel de ruido diferente, su desempeño será pobre. Para evitar esto se procura utilizar redes sencillas limitando también el número de iteraciones del algoritmo de entrenamiento. En la validación de los modelos de redes neuronales se utilizó validación cruzada 10-Fold, debido al poco número de datos disponibles en cada modelo.

Fase 4: Análisis del rendimiento del modelo.

Finalmente, se hace un recuento de los algoritmos de predicción presentes en la literatura y sus errores de predicción. También, se programa el algoritmo de persistencia y el de regresión lineal múltiple, que son utilizados en la literatura como referencia para comparar, y se hallan los errores para ellos. Así, se tiene un marco de referencia con el que se puede comparar el rendimiento de los modelos de predicción. Todo esto con el fin de contrastar los resultados obtenidos en los modelos de predicción para los datos divididos en clústeres, con los resultados del modelo de predicción hecho con el conjunto completo de datos con redes neuronales, regresión lineal y el modelo de persistencia.

En la **Tabla 3** se encuentra un resumen de las actividades asociadas a cada una de las fases del proyecto, o a cada uno de los objetivos específicos.

Tabla 3 Actividades relacionadas con cada una de las fases u objetivos específicos del proyecto.

Objetivo específico 1
Actividad 1: Realizar una revisión bibliográfica de los modelos de predicción de potencia y de las variables estudiadas frecuentemente que tienen efecto sobre la generación de energía fotovoltaica.
Actividad 2: Corregir los datos faltantes para cada variable.
Actividad 3: Elegir las variables a utilizar en el modelo de predicción con un análisis de correlaciones.
Actividad 4: Establecer método de agregación de variables al modelo.
Objetivo específico 2
Actividad 5: Escalar los datos de forma que tengan media cero y desviación estándar 1.
Actividad 6: Calcular el número óptimo de clústeres.
Actividad 7: Ejecutar el algoritmo de agrupación.
Actividad 8: Interpretar los resultados de los clústeres.
Actividad 9: Dividir los datos en el número óptimo k de grupos.
Objetivo específico 3
Actividad 10: Elegir el algoritmo de optimización para los parámetros de la red neuronal.

Actividad 11: Establecer el método de validación de los modelos desarrollados con redes neuronales.
Actividad 12: Entrenar varias redes neuronales alterando varios hiperparámetros para cada clúster.
Actividad 13: Entrenar varias redes neuronales alterando varios hiperparámetros para el conjunto total de datos.
Actividad 14: Calcular el nRMSE para el conjunto de datos de validación de las redes entrenadas.
Objetivo específico 4
Actividad 15: Estimar el error del modelo de persistencia.
Actividad 16: Calcular los parámetros del modelo de regresión lineal multivariada para la potencia.
Actividad 17: Comparar los errores de los modelos desarrollados.
Actividad 18: Redacción informe final.

1.6. Estructura del documento

El contenido de este proyecto se encuentra dividido en seis capítulos de la siguiente forma:

El primer capítulo contiene la descripción del proyecto, en donde se encuentran el marco conceptual, la revisión bibliográfica, el planteamiento del problema, los objetivos y la metodología. Como conclusiones de la revisión bibliográfica, se encuentran las brechas tecnológicas que permiten ver la relevancia de este trabajo. De igual forma, la revisión de la literatura sirve de base para la elección de las variables que se incluyeron en el modelo de predicción y la estructura de éste.

En el segundo capítulo se enuncia el proceso detallado de tratamiento de datos. Detallando el proceso de filtrado utilizando técnicas de agrupación, el tratamiento de los datos faltantes, y el cambio de frecuencia de medición. Así como la elección de las variables del modelo y su método de agregación.

En el tercer capítulo se agrupan los datos resultantes utilizando el algoritmo de agrupación k-means. En este capítulo se explican los pasos del algoritmo k-means y la formulación matemática que los sustenta. De igual forma, se utilizan varios métodos para calcular el número óptimo de clústeres. Posteriormente se dividen los datos en distintos conjuntos y se interpretan los resultados del algoritmo de agrupación

El cuarto capítulo en cambio se explica el proceso matemático de entrenamiento de una red neuronal y se entrenan los modelos para cada uno de los clústeres. Este mismo procedimiento se hace para el conjunto de datos que no ha sido agrupado. De igual forma, se explica el proceso de validación del modelo con validación cruzada K-Fold, y el ajuste del número de neuronas por cada capa.

El quinto capítulo se dedica a estudiar el rendimiento de las redes neuronales desarrolladas comparándolas con modelos base de la literatura. Se desarrollan modelos de regresión lineal multivariada y de persistencia para el conjunto de datos que no ha sido agrupado para tener una base de comparación al evaluar el desempeño del modelo propuesto en este documento.

Cabe resaltar en este punto que la forma en la que se encuentra hecha la división del contenido por capítulos corresponde a la ejecución y cumplimiento de cada uno de los objetivos específicos. El primer y segundo capítulo corresponden al primer objetivo específico. El tercer capítulo hace referencia al segundo objetivo

específico. El cuarto capítulo trata el tercer objetivo específico y el quinto capítulo abarca el último objetivo específico.

Finalmente, en el sexto capítulo se encuentran las conclusiones de este trabajo y recomendaciones a trabajos futuros. Seguido, los documentos citados en este trabajo en la sección de bibliografía.

Capítulo 2: Preparación de los datos

2.1. Obtención de los datos

Los datos que se utilizan en este trabajo de investigación corresponden a una plataforma experimental ubicada en la Universidad del Norte, Puerto Colombia. El diagrama de conexiones de la plataforma se encuentra en la **Figura 13**.

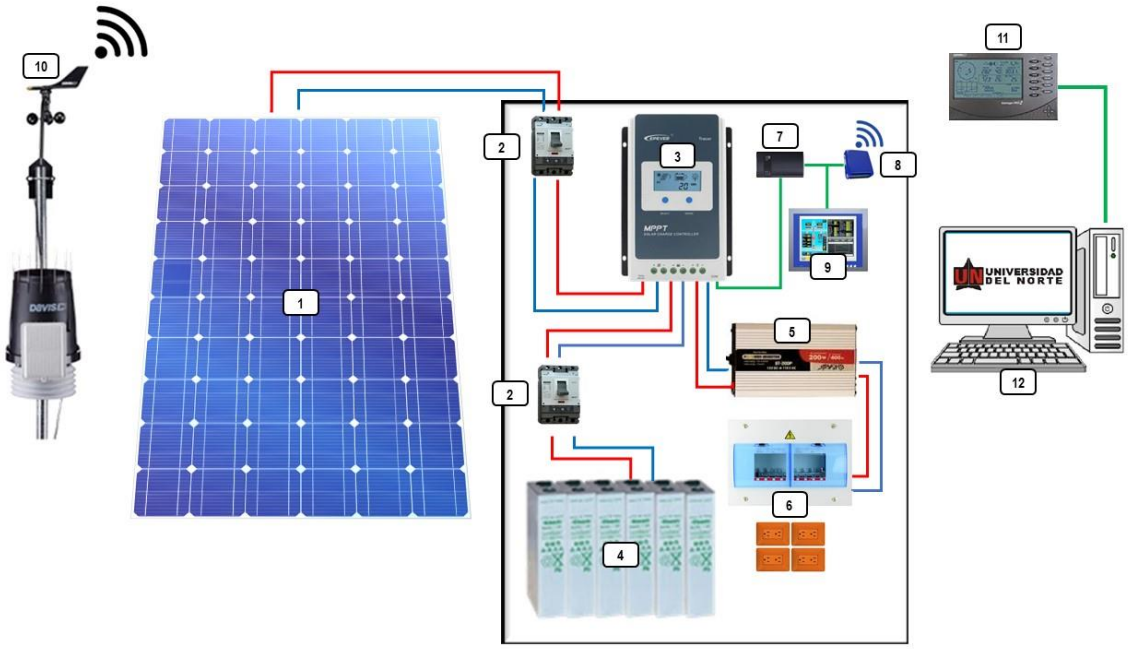

Figura 13 Esquema de conexión de los equipos en la plataforma experimental.

El ángulo de inclinación del panel solar (representado en la **Figura 13** con el número 1) es de o (cero) grados, es decir, horizontal. El lado largo de los paneles está orientado con la línea Norte-Sur. La descripción de los ítems en la **Figura 13** se encuentra en la **Tabla 4**.

Tabla 4 Descripción de los elementos en la estación experimental.

Número del ítem	Descripción
1	Panel solar.
2	Interruptores de mantenimiento.

3	Controlador de carga.
4	Baterías.
5	Inversor DC/AC.
6	Tablero de distribución.
7	Almacenador de datos.
8	Comunicador wifi.
9	Interfaz hombre máquina del sistema de adquisición de datos eléctricos.
10	Estación meteorológica.
11	Interfaz hombre máquina de la estación meteorológica.
12	Computador.

En la **Tabla 5** se encuentran las especificaciones del panel del que se obtuvieron los datos experimentales.

Tabla 5 Especificaciones técnicas del panel solar en condiciones estándar.

Panel solar	
Tipo de panel	Policristalino, 6”
Referencia	CS6X 310P
Número de celdas	72 (6 x 12)
Potencia nominal	310 W
Voltaje de circuito abierto	44.9 V
Corriente de corto circuito	9.08 A
Eficiencia del panel	16.16%
Dimensiones del panel	1954 x 982 x 40 mm
Peso	22 kg
Área	1.92 m^2

De esta plataforma experimental se obtuvo la potencia instantánea del panel con un tiempo entre mediciones de 11 min. Un sensor de temperatura se hallaba localizado en el centro del panel solar, permitiendo tener valores de temperatura cada minuto. De igual forma, a pocos metros del panel solar se encontraba una estación meteorológica realizando mediciones cada 10 minutos. Las variables consideradas en este estudio se encuentran en la siguiente tabla.

Tabla 6 Frecuencia de medición y nombre de las variables utilizadas en el modelo de predicción de potencia.

Variable	**Nombre**	**Frecuencia de medición**
Temperatura ambiente	Ta	Cada 10 min.
Humedad relativa	HR	Cada 10 min.
Velocidad del viento	VV	Cada 10 min.
Irradiancia en el plano horizontal	G	Cada 10 min.
Índice UV	UV	Cada 10 min.
Temperatura del panel	Tc	Cada 1 min.

Potencia instantánea	P	Cada 11 min.

Las celdas en el panel solar están ubicadas en una malla de 6 celdas de ancho y 12 de largo. La temperatura del panel fue medida en seis puntos que se encontraron distribuidos en la cara inferior de éste, como se muestra en la **Figura 14**. T_{c_1}, T_{c_2}, T_{c_3}, T_{c_4}, T_{c_5}, T_{c_6} son las temperaturas de los puntos 1, 2, 3, 4, 5 y 6 respectivamente.

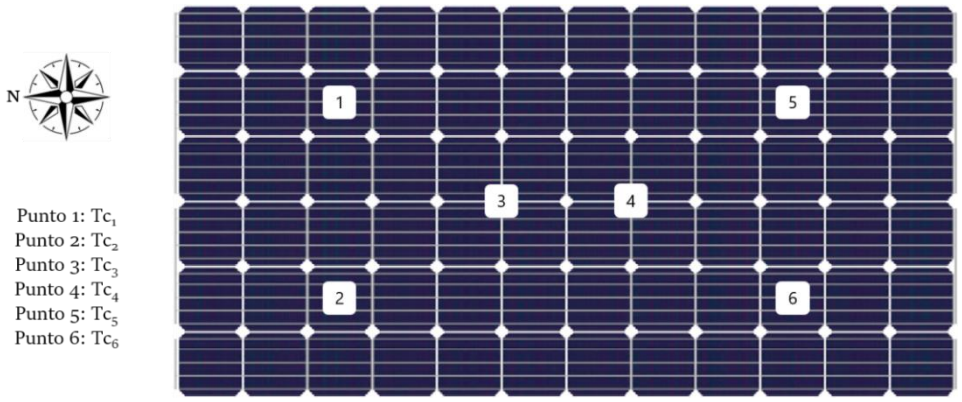

Figura 14 Posición de las termocuplas en el panel solar. Las termocuplas se encuentran en la parte de posterior del panel.

En la **Figura 15** se muestra el promedio de cada una de las temperaturas del panel solar. Gráficamente se puede ver que los valores de temperatura en cada punto del panel en el que se midió son muy parecidos. Por ello, en este trabajo se utiliza un promedio aritmético de los distintos valores medidos para cada instante.

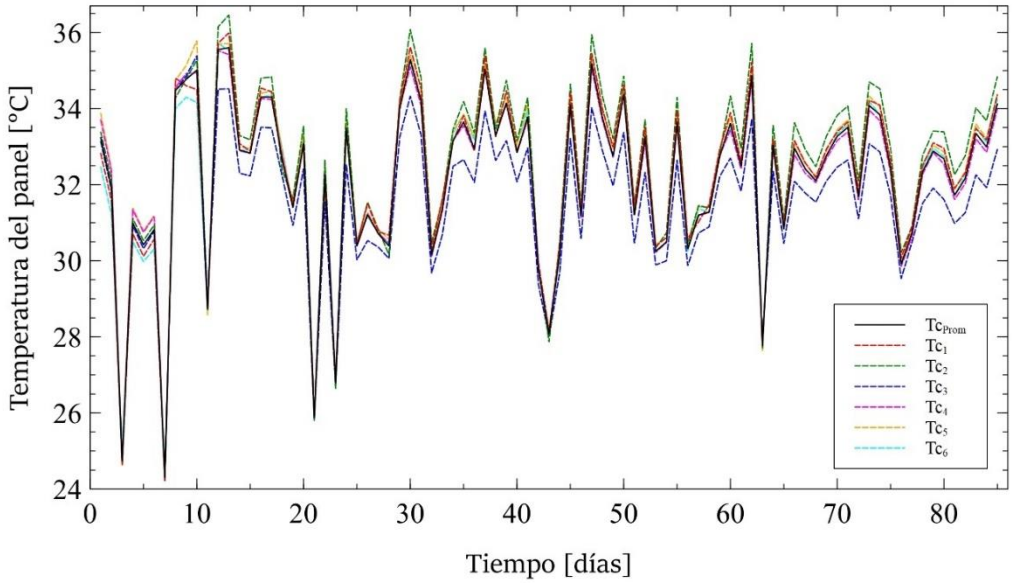

Figura 15 Promedios de las temperaturas

En la mayoría de los casos, antes de poder entrenar un modelo, es necesario hacerle un tratamiento a los datos para que sean aptos para el modelo (Loy, 2019). Estos tratamientos pueden incluir codificación de

variables categóricas o manejo de datos faltantes. En este caso, todas las variables que se miden son numéricas, pero hay algunos datos que pueden no aportar gran información al modelo, o que tienen observaciones faltantes. En la siguiente sección se detalla el proceso que se aplica en los datos antes de entrenar los modelos de predicción de potencia fotovoltaica.

2.2. Tratamiento de los datos

El tratamiento de los datos se llevó a cabo con el software estadístico R y su interfaz de usuario RStudio® por sus ventajas en el manejo de datos, gran cantidad de herramientas estadísticas y su facilidad para la creación de gráficos. En la **Tabla 7** se encuentra una comparación entre este software con otros softwares estadísticos.

Tabla 7 Comparación de varios softwares estadísticos (R, SAS, Stata, SPSS). Adaptado de (Dinov, 2018).

Software	Ventajas	Desventajas
R	R es un software activo actualmente (developers, packages). Tiene excelente conectividad con otros tipos de datos y sistemas. Es versátil para resolver problemas en cualquier dominio. Es gratis y de código abierto. Extensibilidad: R admite extensiones, por ejemplo, para manipulación de datos, modelado estadístico y gráficos. Una gran comunidad activa y comprometida apoya a R. Sitios web de preguntas y respuestas (Q&A) incomparables. R se conecta con otros lenguajes (Java/C/JavaScript/Python/Fortran) y sistemas de bases de datos, y otros programas, SAS, SPSS, etc.	Mayormente lenguaje de script. Curva de aprendizaje inclinada. Trabaja los procesos en la RAM.
SAS	Maneja conjuntos de datos grandes. Es comúnmente usado en negocios y aplicaciones gubernamentales.	Es costoso. El lenguaje de programación es un poco anticuado.
Stata	Análisis estadísticos sencillos.	Solo maneja estadística clásica.
SPSS	Apropiado para principiantes, interfaz sencilla.	Tiene debilidad en los últimos algoritmos más avanzados. Le falta robustez en los métodos estadísticos y los de encuestas.

Las mediciones utilizadas en esta investigación fueron tomadas de forma no continua en el tiempo entre el 10 de abril del 2019 hasta el 22 de noviembre de ese mismo año. En total se cuenta con 85 días de datos, que comprenden los días 6, 7 y del 21 al 25 de febrero, del 10 al 12 de abril y del 9 de septiembre al 22 de noviembre del 2019. Sin embargo, varios de estos días representan valores atípicos que aumentan el ruido de los datos, que bajan el desempeño del algoritmo de agrupación y dificultan el ajuste de los modelos de predicción. Por ello, el primer filtro aplicado a los datos fue eliminar todos aquellos días donde el funcionamiento del panel solar no fue adecuado. Estos fueron días donde a pesar de tener valores altos de irradiancia, la potencia promedio en el día fue casi cero. Una vez eliminados estos días, el resultado fue 69 días de mediciones. En la **Tabla 8** se presenta un resumen del promedio diario para estos estas 69 muestras.

Tabla 8 Resumen de las variables medidas después de eliminar los valores promedio por día de potencia cercanos a cero. En esta tabla se encuentran promedios diarios.

Resumen	G [W/m²]	UV	Tc [°C]	Ta [°C]	HR [%]	VV [m/s]	P [W]
Mínimo	55.11	0.3804	24.30	25.26	75.78	0.3403	10.25
1er cuartil	145.20	0.9869	30.72	27.69	77.45	0.9056	34.28
Mediana	183.55	1.4410	32.48	28.59	78.43	1.0909	44.44
Promedio	179.13	1.1944	31.97	28.37	78.45	1.2829	43.31
3er cuartil	217.85	1,4156	33.27	29.15	79.28	1.4608	53.80
Máximo	269.79	1.8455	35.60	30.60	81.77	3.8653	75.69

Después de haber aplicado este primer filtro en la potencia, se siguieron encontrando datos por fuera de la línea de tendencia como se muestra en **Figura 16**. Estos datos atípicos producen ruido que posteriormente aumentarán el error del modelo de predicción, por ello se continúa con los filtros hasta eliminarlos.

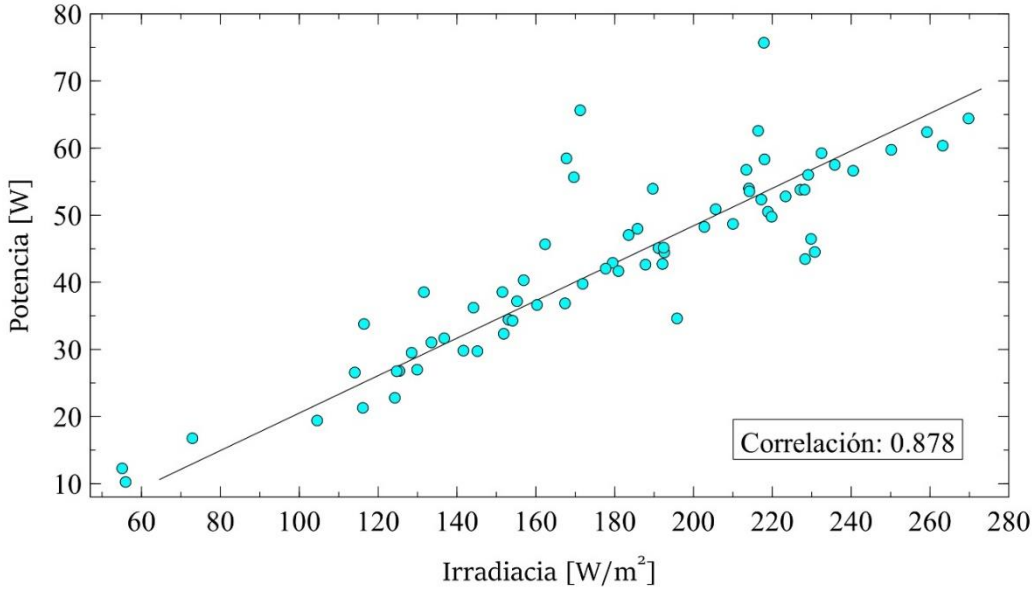

Figura 16 Potencia vs. Irradiancia. Gráfica obtenida con los promedios diarios de potencia e irradiancia después de haber eliminado los días donde la potencia era casi cero.

Los valores atípicos fueron eliminados buscando tener una mayor correlación de la potencia con la irradiancia, que es la variable que por naturaleza tienen mayor efecto en la producción de energía fotovoltaica. Estos valores atípicos consistieron mayormente de días donde la potencia fue cercana a cero, siendo muy baja para la irradiancia que hubo en promedio en el día.

Para eliminar estos datos por fuera de la línea de tendencia se utilizó el algoritmo de agrupación k-means. Éste utiliza la distancia euclidiana para hacer clústeres. El objeto de utilizar este método es encontrar aquel grupo de datos donde las mediciones de potencia son muy bajas, o no están relacionados con los valores de irradiancia. Explicar el algoritmo de agrupación no es el propósito de este capítulo. En cambio, el detalle de la sustentación matemática y los pasos de este algoritmo se detallan en el **Capítulo 3**, por ahora, solo se

muestran los resultados obtenidos. Los valores utilizados fueron nuevamente promedios diarios de los datos. El resultado de la agrupación con k-means fue conseguir tres clústeres descritos en la **Tabla 10**. Al analizar el contenido de los clústeres encontramos lo siguiente, el primer clúster tiene 39 días, el segundo tiene 23 y el tercero tiene 7. En el primero y el segundo hay una alta correlación positiva entre la potencia fotovoltaica y la irradiancia solar, como es de esperarse. En cambio, en el tercero la correlación es negativa, indicando invalidez de los valores de potencia para esos días. El tercer clúster es aquel en donde el funcionamiento del panel solar no corresponde al adecuado. Esto se ve en la correlación de la potencia con la irradiancia y temperatura de la celda. Para el caso de la irradiancia, la correlación es negativa. Y para el caso de la temperatura de la celda, esta correlación es casi igual a cero. Por ello se eliminaron también estos datos del conjunto.

Tabla 9 Correlación entre P y G y entre P y Tc para cada clúster.

Clúster	Número de días	Correlación entre P y G	Correlación entre P y T_c
Clúster 1	39	0.713	0.188
Clúster 2	23	0.932	0.802
Clúster 3	7	-0.575	-0.001

Adicional a los datos que se eliminaron, se descartaron también aquellos días incompletos, dejando como resultado 57 días de medición que serán utilizados en los modelos de predicción. Luego de eliminar todos aquellos días que solo agregarían ruido al modelo de predicción se obtuvo una correlación entre la potencia y la irradiancia es de 0.965.

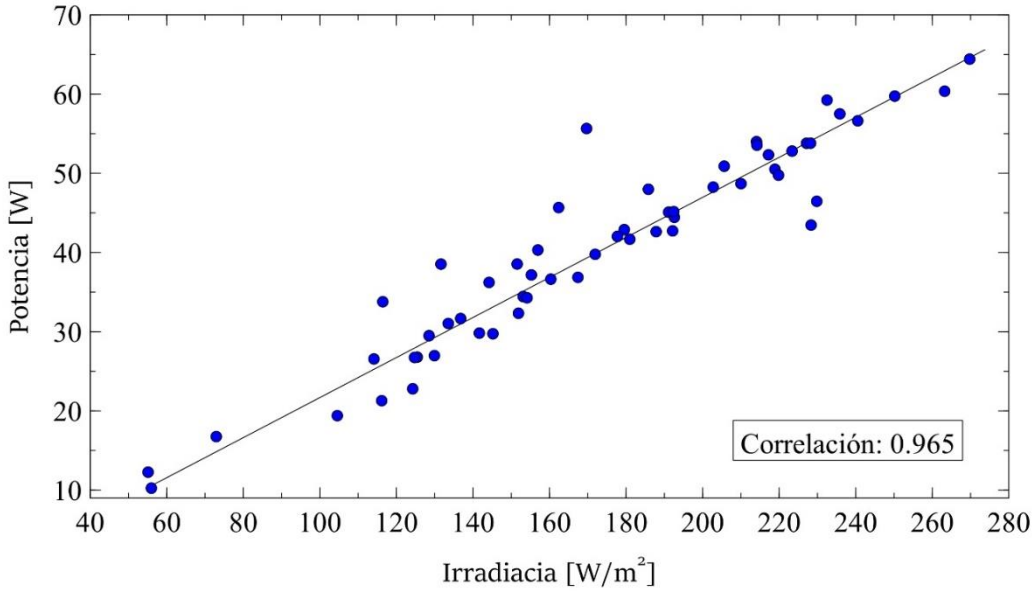

Figura 17 Potencia vs. Irradiancia. Gráfica obtenida con los promedios diarios de potencia e irradiancia después de haber filtrado datos ruidosos con agrupación por k-means.

Esta mayor correlación en los datos diarios también puede ser observada al graficar los valores de potencia promedio por día con los valores de irradiancia promedio en esos días. Esta gráfica se encuentra en la **Figura 18** y muestra la similitud entre estas dos variables.

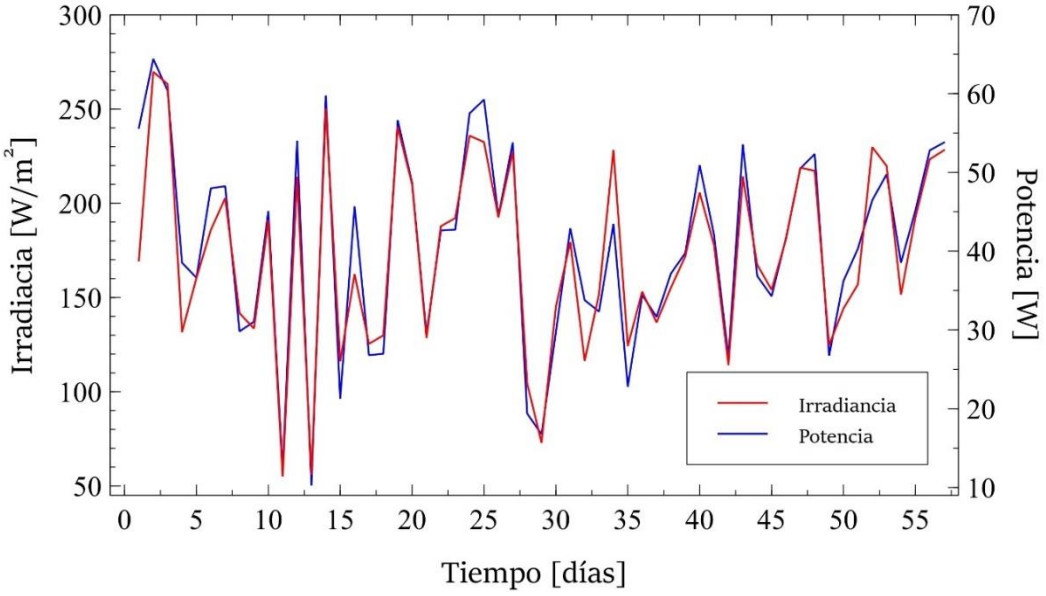

Figura 18 Irradiancia y potencia diarias en los distintos días. Gráfica obtenida con los promedios diarios de potencia e irradiancia después de haber filtrado datos ruidosos con agrupación por k-means.

2.2.1. Imputación de datos faltantes

Cuando a algunas variables les hacen falta valores, la forma más sencilla de solucionar el problema es eliminando la fila en la que se encuentra este valor. Sin embargo, esta solución tiene como desventaja la pérdida de información que puede ser relevante. Aún más en casos donde la información es escasa. Otra solución al problema de los valores faltantes es imputar datos. Existen varias formas de hacer la imputación, una de ellas es reemplazar los datos faltantes por el promedio total de los datos. También se puede reemplazar por la moda. Estas soluciones no son las más convenientes en muchos casos, sobre todo cuando el rango de la variable es amplio.

Para los datos de esta investigación, se cuenta con algunos datos faltantes en todas las variables. En la **Figura 19** se muestra la distribución de los datos faltantes en la temperatura del panel que se mide cada minuto. Como se mencionó anteriormente, cada variable se mide a una frecuencia diferente, y es necesario, antes de entrenar el modelo, llevar estas frecuencias de medición a una sola. Este procedimiento se mencionará más adelante.

Con respecto a la temperatura de la celda, se cuenta con 82080 datos, de los cuales 105 son valores faltantes o NAs (dato faltante). Esto representa el 0.128% de los datos. Como regla general se tiene que, si menos del 5% de los datos son NAs, es aceptable hacer imputación de datos (Cambridge Spark, 2019). Para el caso de la temperatura de la celda, los datos faltantes se encuentran espaciados, es decir, no falta un día entero de datos. Por el contrario, los NAs no son consecutivos. Como la temperatura es una variable que no varía con mucha rapidez, se puede hacer interpolación lineal para hallar el valor de estos datos faltantes.

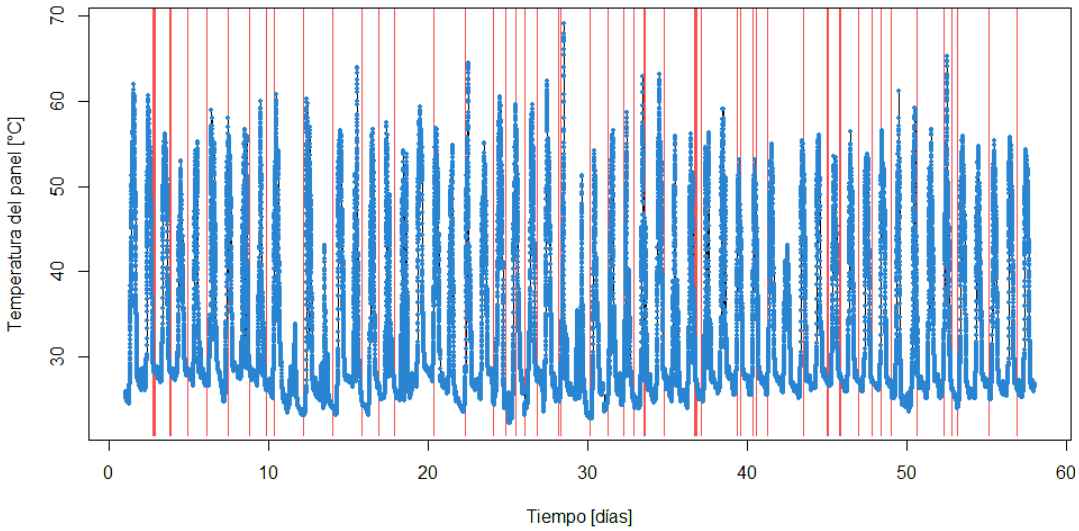

Figura 19 Distribución de los datos faltantes en la temperatura del panel visualizadas como líneas rojas verticales. El grosor de la línea indica la cantidad de datos faltantes

Este mismo procedimiento se hace para las demás variables, rectificando que los valores faltantes no sean consecutivos por tiempos largos y que el número de NAs no represente más del 5% del total de los datos. En la **Tabla 10** se encuentra el promedio de cada variable antes y después de hacer la imputación. Es posible apreciar que la diferencia entre los valores es muy pequeña.

Tabla 10 Cambio de los promedios de las variables después de imputar los datos faltantes.

Variable	Promedio antes de la imputación	Promedio después de la imputación
P [W]	40.67	40.77
G [W/m2]	172.7	172.5
UV	1.210	1.206
Tc [°C]	32.24	32.24
Ta [°C]	28.53	28.53
HR [%]	78.54	78.54
VV [m/s]	1.110	1.110

Al haber solucionado todos los NAs, se estudia el efecto de las variables ambientales y temperatura de la celda sobre la potencia de los paneles solares. La mayoría de las variables tienen una relación numérica alta con la potencia de los paneles, siendo esto un indicio de que es buena idea utilizar estas variables en el modelo de predicción. Cabe aclarar que una correlación alta no implica necesariamente efecto de una variable sobre la otra, pero en este caso se sabe de estudios previos y de la física detrás del fenómeno fotovoltaico que las variables elegidas en el modelo tienen relación con la potencia del panel. La variable cuya correlación con la potencia es más alta es la irradiancia como es de esperarse. A pesar de que mayor temperatura de la celda implica menor eficiencia del panel, mayor irradiancia se ve reflejada en mayor potencia y temperatura de la celda. Por otro lado, la velocidad del viento es la variable que está menos

correlacionada con la variable de salida. Esto implica que su aporte al modelo de predicción puede ser poco, o puede ser negativo, ya que aportara solo ruido.

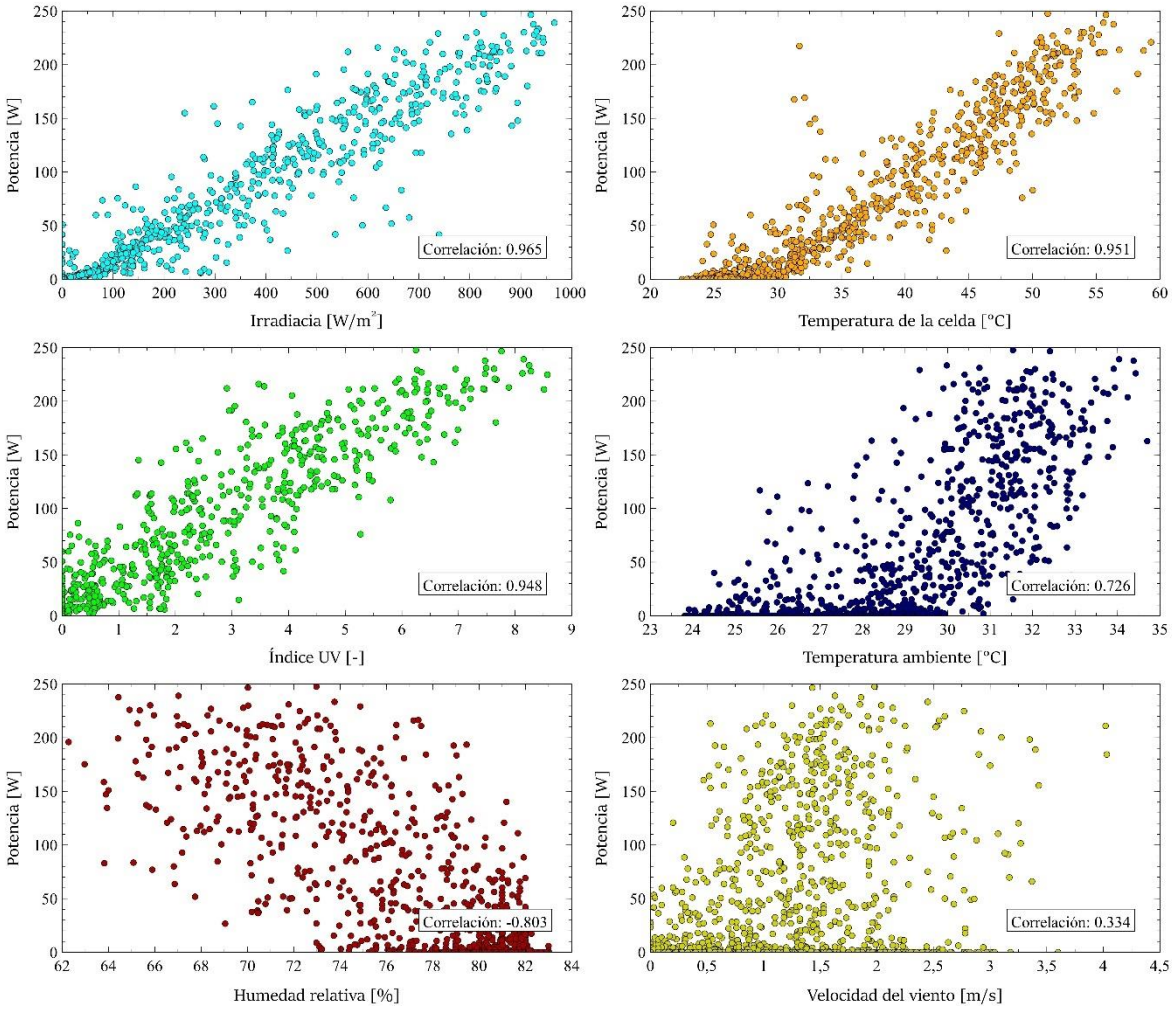

Figura 20 Matriz con gráficos de dispersión y correlaciones de la potencia contra las variables de entrada al modelo.

2.3. Método de agregación de variables

Cada una de las variables se mide con una frecuencia de muestreo distinta. Es necesario llevarlas todas a la misma frecuencia antes de poder entrenar el algoritmo de predicción. En la **Tabla 11** se encuentra el método de agregación para las variables de entrada al modelo de predicción de potencia de los paneles. Al hacer promedios horarios se incluyen valores de las variables durante la noche, donde la irradiancia es cero y no hay generación de energía. La **Figura 20** utiliza los promedios de las variables por hora. Para hallar estos promedios, en el caso de la temperatura de la celda, donde las mediciones fueron hechas cada minuto, se hicieron promedios de los 60 datos en cada hora. Para las variables ambientales, las mediciones fueron hechas cada 10 min. Por ello, se tomaron los 6 datos correspondientes a cada hora y se promediaron. Para la potencia, las mediciones se encontraban hechas cada 11 minutos, lo que dificulta los promedios. Así que

se llevó la frecuencia de muestreo de la potencia a cada 1 minuto, por medio de interpolación lineal y después, se promediaron los 60 datos correspondientes a cada hora.

Tabla 11 Método de agregación de variables al modelo de predicción de potencia.

Variable	Tipo de medición	Cálculo
P	Directa	Promedio simple horario
G	Directa	Promedio simple horario
UV	Directa	Promedio simple horario
Tc	Directa	Promedio simple horario
Ta	Directa	Promedio simple horario
HR	Directa	Promedio simple horario
VV	Directa	Promedio simple horario

La **Figura 20** muestra una matriz con los gráficos de dispersión de las variables y la correlación entre la potencia y cada variable independiente. Estas correlaciones dan una idea clara de las variables a incluir en el modelo de predicción de potencia. Las variables que tienen mayor efecto sobre la generación de potencia fotovoltaica son la irradiancia y la temperatura de la celda, con correlaciones de 0.965 y 0.951, respectivamente. Sin embargo, cuando no se tienen mediciones de temperatura del panel, se suelen usar las de temperatura ambiente, que en este caso tiene una correlación de 0.726. La velocidad del viento es una variable que tiene efecto en la temperatura del panel (Dolara, Leva, & Manzolini, 2015). A mayor velocidad del viento, menor temperatura del panel. A su vez, la temperatura del panel disminuye la eficiencia de los paneles solares cuando aumenta en magnitud. De hecho, la velocidad del viento se tiene en cuenta en varias correlaciones para la eficiencia de los paneles solares (Skoplaki & Palyvos, 2009). A pesar de que existe una relación indirecta o de segundo orden entre la velocidad del viento y la temperatura de los paneles solares (Fuentes, Nofuentes, Aguilera, Talavera, & Castro, 2007), la correlación entre la velocidad del viento y la potencia es baja, de 0.334, por lo que incluirla en el modelo de regresión, aumentaría el error de validación. El efecto de esta variable en cada clúster se estudia en el siguiente capítulo. La humedad relativa es otra variable que se suele usar en los modelos de predicción de potencia (Isha, Chaudhary, & Chaturvedi, 2020). Aunque esta última y el índice UV son variables que aparecen con mucha más frecuencia en los modelos de degradación de potencia de los paneles solares fotovoltaicos (Fernandes, Torres, Morgado, & Morgado, 2016; N. C. Park, Oh, & Kim, 2013; N. Park, Kim, Kim, & Moon, 2017). Por esta razón se han incluido en este estudio, además de ser parte de las condiciones ambientales del lugar en el que se encuentra el panel y encontrarse altamente correlacionadas con su potencia.

Capítulo 3: Agrupación de los datos

Uno de los resultados de la revisión bibliográfica fue conocer que la mayoría de los documentos académicos donde se hace predicción de potencia tienen un proceso de clasificación de los datos antes de ejecutar el algoritmo de predicción (Cheng, Guo, Wang, & Zafar, 2017; Pulipaka & Kumar, 2016). Es decir, combinan distintas técnicas para hacer el modelo de predicción. Entre estas, una de las más comunes es la clasificación con kNN (*k nearest neighbors*). En este trabajo se hizo clasificación de los datos con el algoritmo de k-means antes de entrenar las redes neuronales y conseguir distintos grupos de datos con menor variación de los datos dentro de ellos. El objetivo de esto es conseguir grupos donde los valores de la potencia sean similares y así mejorar el rendimiento del modelo de predicción (Raza & Khosravi, 2015).

3.1. Algoritmo de agrupación k-means

El algoritmo k-means es un algoritmo de aprendizaje no supervisado. Esto implica que no es necesario indicar cuáles datos son de entrada y cuáles de salida. En este algoritmo, k es el número de conjuntos en los cuales se agruparán los datos. Este método se basa en ubicar cada par de datos en el conjunto o clúster más cercano. Un clúster se define como un grupo donde todos los objetos son similares entre sí, y diferentes a los objetos en otro clúster. El algoritmo k-means es muy común debido a su sencillez y robustez. Su ejecución se encuentra basada en pocos pasos sencillos que llevan a agrupar los datos de una forma eficiente (Hartigan & Wong, 1979). Entre ellos está primero definir el número de clústeres, k, y ubicar k centroides aleatorios para esos clústeres. Luego asignarle a cada centroide los datos más cercanos a este. Y finalmente recalcular los centroides con el promedio de cada clúster. Estos pasos se repiten hasta que cada centroide no cambie de posición. Lo anterior se puede ver con más claridad en la **Figura 21**.

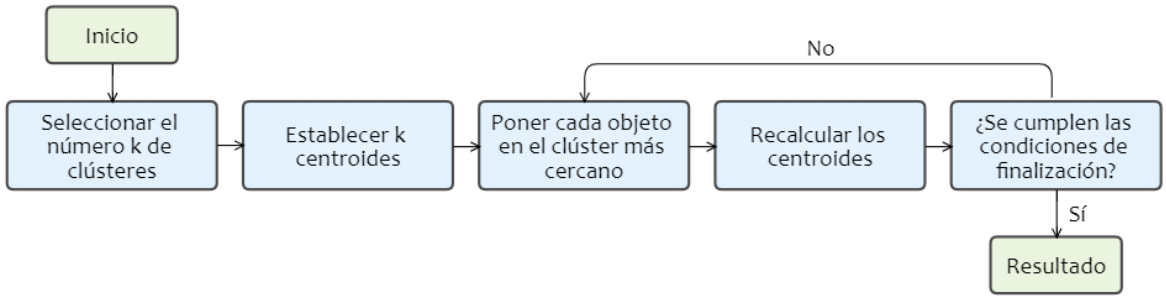

Figura 21 Diagrama de flujo con los pasos del algoritmo de agrupación k-means.

Una opción para generar los grupos de datos era entrenar todos los datos promediados por hora, sin embargo, de esa forma el algoritmo de agrupación entrega dos clústeres. En uno de ellos se encuentran las horas del día donde la potencia es alta y en el otro los valores de potencia baja, en la noche y horas tempranas de la mañana. Más allá de dividir cada día en dos grupos cuyos límites horarios son prácticamente iguales entre un día y otro, se procedió a clasificar los datos en aquellos donde la irradiancia fuera alta, o día soleado, y aquellos donde la irradiancia es baja, o día nuboso. Por ello se ejecutó el algoritmo de k-means en los datos con promedios diarios.

Para ubicar los datos en el clúster más cercano se requiere un método que permita medir la disimilitud entre dos puntos dados. En este caso, se utiliza la distancia euclidiana entre dos puntos. En la **Ec. 10** se muestra la distancia euclidiana entre dos puntos.

$$d_{euc} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \textbf{Ec. 10}$$

Donde n es el número de observaciones y x e y son los dos elementos cuya distancia quiere ser calculada. Básicamente la agrupación de k-means consiste en definir grupos para que se minimice la variación total dentro de cada clúster. Esta variación también es conocida como variación total dentro del clúster. Hay varias formas del algoritmo de agrupación k-means, sin embargo, el algoritmo estándar es el algoritmo Hartigan-Wong (Hartigan & Wong, 1979). En este se define la variación dentro del clúster como la suma de las distancias al cuadrado de las distancias euclidianas entre los elementos y el centroide correspondiente. Esto puede ser calculado como se describe a continuación.

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad \textbf{Ec. 11}$$

Donde x_i es un punto perteneciente al clúster C_k, y μ_k es el valor promedio de los puntos pertenecientes a C_k. La variación total dentro de los clústeres (WSS) se expresa como,

$$WSS = \sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2 \qquad \textbf{Ec. 12}$$

Esta variación total dentro de los clústeres es una medida de qué tan bien se ajustan los datos dentro de ellos. Al ser un valor pequeño indica que los datos dentro de los clústeres están “bien ajustados”.

3.1.1. Método del codo

Una de las desventajas del algoritmo k-means es el hecho de tener que entregar un número de clústeres como parámetro de entrada antes de conocer cuál es el mejor número de clústeres. Sin embargo, existen métodos que ayudan en la elección del valor óptimo de k. Uno de ellos es el método del codo (elbow method), que consiste en minimizar WSS de la **Ec. 12**. El procedimiento en el que se aplica consiste en ejecutar el algoritmo k-means para distintos valores de k, por ejemplo, de k=2 hasta k=10. Para cada valor de k se calcula la variación total dentro de los clústeres y se grafica esta versus k. Se sabe que mientras más clústeres se calculen para un conjunto de datos, menor será esta variación total, pues más cercanos estarán

los puntos entre sí dentro de cada clúster. Sin embargo, generalmente se le considera un valor óptimo de k a aquel valor a partir del cual la reducción en WSS ya no es tan notoria (Boehmke, 2019).

Dado que este método funciona calculando distancias, es necesario escalar los datos antes de ejecutarlo. La forma en la que se escalaron los datos fue primero centrándolos al sustraer la media en cada variable. Luego se dividieron entre la desviación estándar.

Las 57 observaciones escaladas se utilizaron para ejecutar el algoritmo de k-means, utilizando primeramente el método del codo para hallar el número óptimo de clústeres. Este se muestra en la siguiente figura.

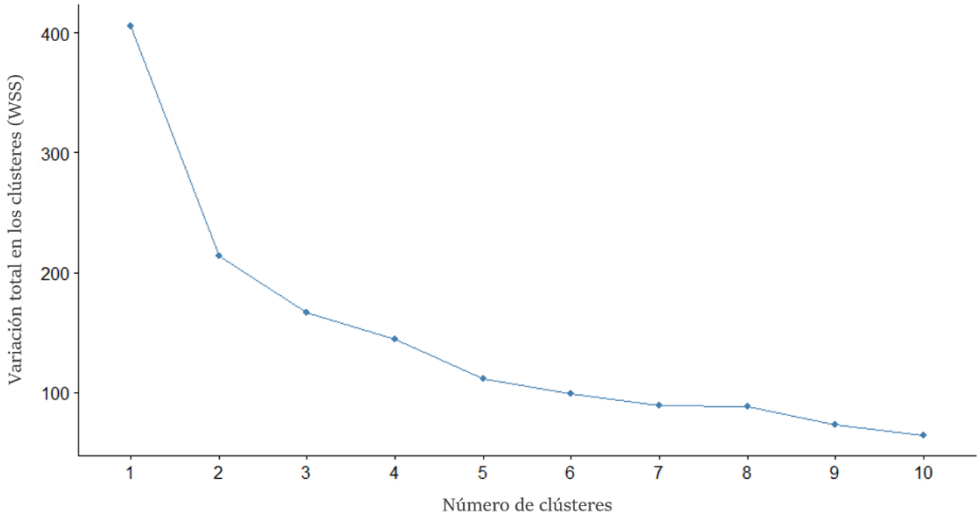

Figura 22 Método del codo para determinar el número óptimo de clústeres. Gráfica obtenida con los promedios diarios de potencia e irradiancia.

La **Figura 22** cuenta con la forma decreciente esperada, y se alcanza a ver con que el codo o punto de inflexión es para dos clústeres. A partir de dos clústeres la reducción en la variación total dentro de los clústeres no es tan grande.

3.1.2. Método de silhouette

Sin embargo, se puede utilizar un segundo método que verifique los resultados obtenidos por el primero. Este es el método de silhouette, que básicamente consiste en medir la calidad de la agrupación. Dicho de otra forma, mide qué tan bien un dato se parece a los demás en su clúster. Un alto valor promedio de silhouette indica que los datos pertenecen bien a los clústeres a los cuales fueron asignados.

Para calcular el promedio de silhouette, se mide primero la distancia promedio entre un punto i y los demás puntos pertenecientes al clúster en el que se encuentra, en este caso representado como C_i en la **Ec. 13**.

$$a(i) = \frac{1}{|C_i| - 1} \cdot \sum_{j \in C_i, j \neq i} d(i,j) \qquad \textbf{Ec. 13}$$

En la **Ec. 13**, el término $d(i,j)$ representa la distancia euclidiana entre los puntos i y j, tal y como se muestra en **Ec. 10**. Podemos interpretar $a(i)$ como una medida de qué tan bien i está asignado a su grupo. Por ende,

para menores valores de $a(i)$, mejor es la asignación de i en el clúster. Por otro lado, definimos en la **Ec. 14** la disimilitud promedio del punto i con los puntos en los demás clústeres C_k, donde $C_k \neq C_i$.

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j) \qquad \textbf{Ec. 14}$$

$b(i)$ es la mínima distancia promedio entre i y todos los puntos en cualquier otro clúster donde i no pertenece. Un valor grande de $b(i)$ indica que el punto i no encaja bien en los demás clústeres. El valor silhouette promedio para el punto i entonces se define como,

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad \textbf{Ec. 15}$$

Donde $s(i)$ es el valor silhouette promedio para el punto i. El cálculo de la **Ec. 15** se realiza para todo el conjunto de datos. El valor promedio de silhouette en todo el conjunto de datos entonces muestra qué tan bien ha sido agrupado. El resultado del método de silhouette se muestra en la **Figura 23**, donde se puede apreciar que el número óptimo de clústeres es 2.

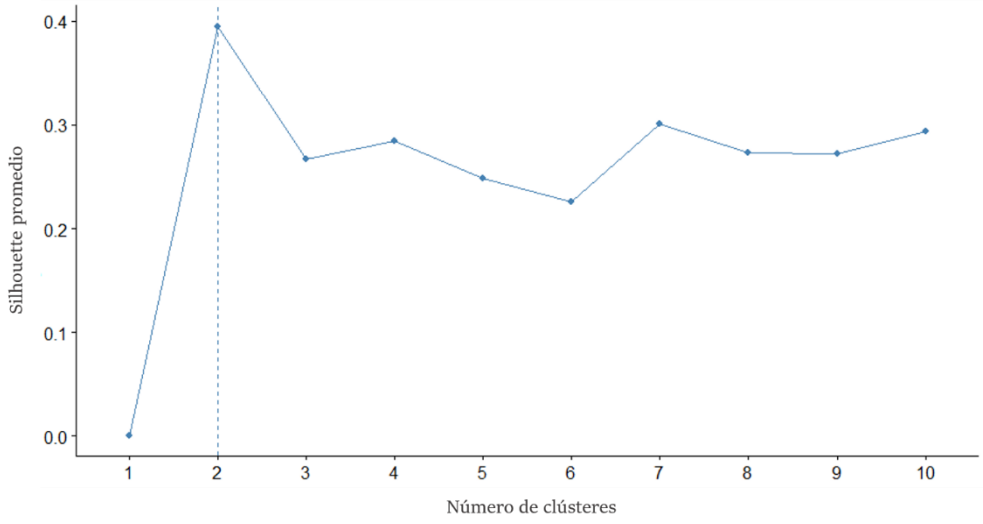

Figura 23 Método de silhouette para determinar el número óptimo de clústeres. Gráfica obtenida con los promedios diarios de potencia e irradiancia.

Tanto en el método del codo (**Figura 22**) como en el de silhouette (**Figura 23**) es notorio que el número óptimo de clústeres es 2. Esto se hace evidente al ver los resultados en los que se obtuvieron 3 clústeres y se eliminó uno, dejando como resultado 2. El clúster 1 tiene 38 días y el clúster 2 tiene 19. Si se tiene en cuenta que los datos para realizar la agrupación fueron promedios diarios, pero en el modelo de predicción se utilizan datos con frecuencia de medición horaria, el número real de datos en el clúster 1 es 912, y en el clúster 2 es de 456 observaciones.

Como se tienen más de tres variables, no es posible hacer una gráfica que permita ver el contenido de los clústeres. Para poderlos visualizar se utiliza el método de componentes principales (PCA), en el que se encuentran las dimensiones que expresan la mayor variabilidad de los datos. En la **Figura 24** se ve esta representación de los clústeres para distintos valores de k. Esta gráfica permite visualizar de mejor manera

el agrupamiento o similitud de las muestras. La dimensión 1 expresa el 71.6% de la variabilidad de los datos y la dimensión 2 el 10.8%. Cabe resaltar que en esta representación gráfica de los datos hay valores de un clúster que se ven más cercanos a otro clúster. Sin embargo, la distancia de los puntos al centroide del clúster se calcula con la **Ec. 10** para las 7 variables que se estudian en este documento, que no se pueden visualizar en una gráfica.

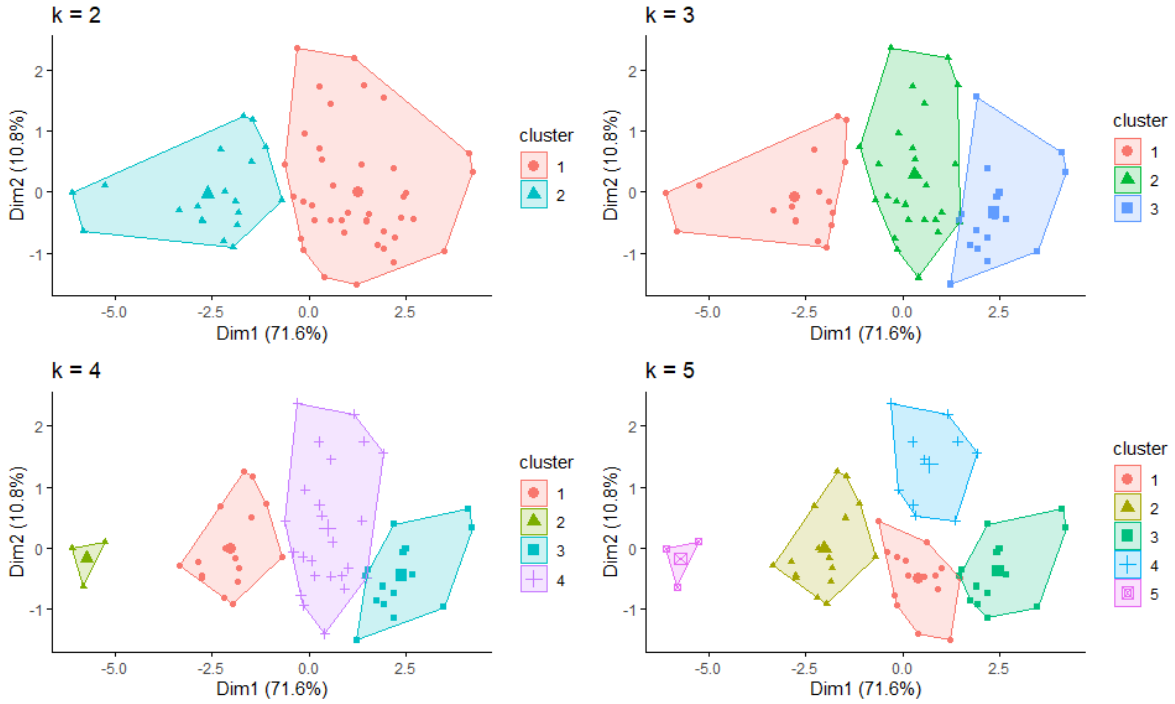

Figura 24 Representación gráfica de los distintos clústeres para valores de k iguales a 2, 3, 4 y 5.

En este punto del documento cabe resaltar la importancia de hacer un tratamiento de los datos antes de ajustar cualquier tipo de modelo. Dependiendo del sistema que se estudie, puede haber muchos datos que solo desmejoren la calidad del modelo de predicción, puesto que no entregan información coherente con los demás datos (Isik, Ozden, & Kuntalp, 2012).

Al dividir los datos en dos grupos, donde cada uno de los grupos tiene información que es homogénea, se puede ver como se sigue cumpliendo la relación entre la potencia y las variables de entrada. Se conservan valores altos de correlación para la potencia fotovoltaica y la irradiancia, temperatura del panel, temperatura ambiente, índice UV y humedad relativa. Es decir, la relación entre variables de entrada y salida para los clústeres es similar a la del conjunto completo de datos. Esto lo vemos al comparar la forma de los diagramas de dispersión y las correlaciones presentes en la **Figura 20**, con la **Figura 25** y la **Figura 26**. Por otro lado, la correlación entre la potencia y la velocidad del viento sigue siendo muy baja en ambos clústeres. Es de 0.314 en el primero y de 0.323 en el segundo. Esta variable falla al explicar el comportamiento de la potencia, razón por la que no se incluye en el modelo de predicción.

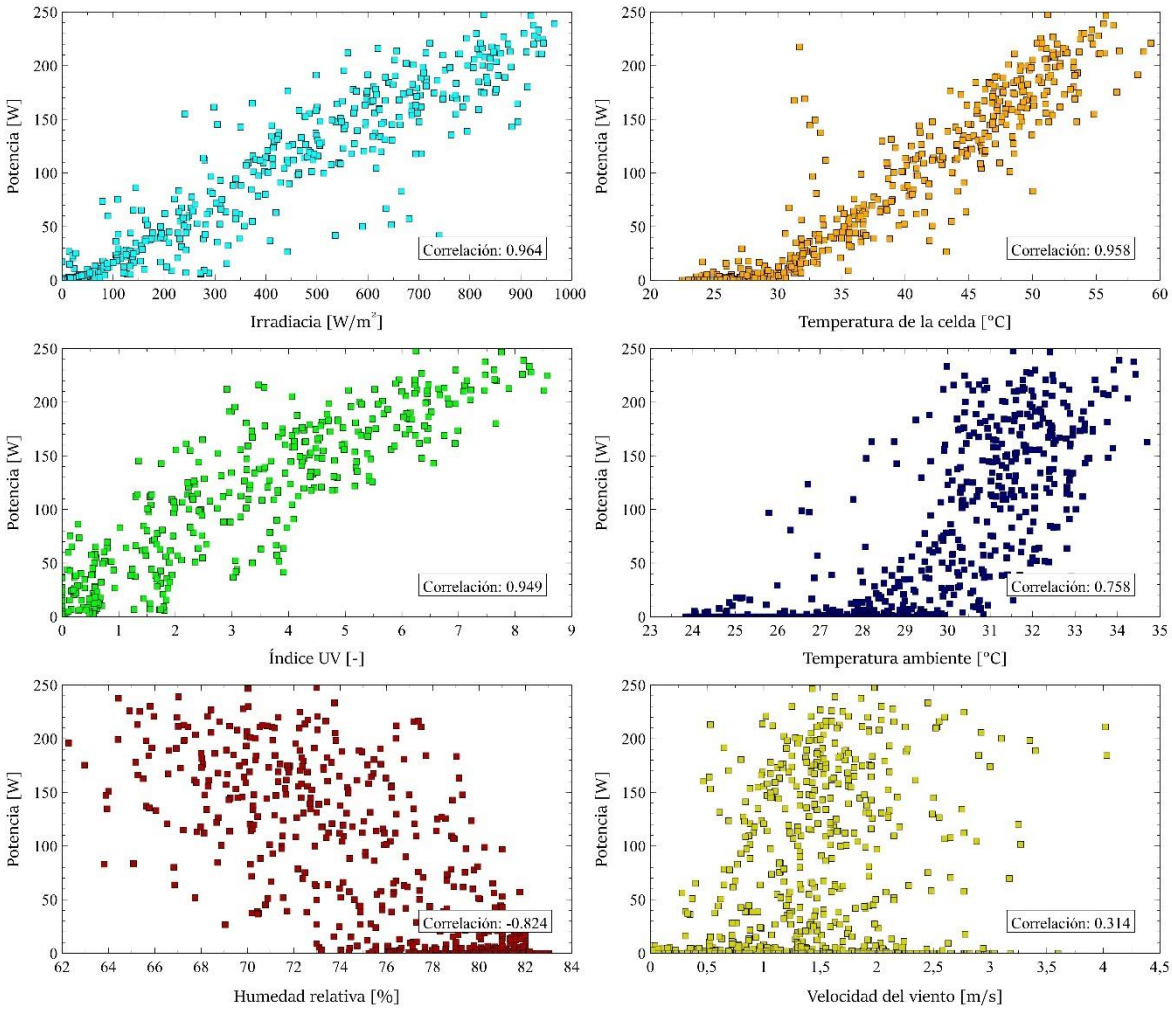

Figura 25 Matriz con gráficos de dispersión y correlaciones de la potencia contra las variables de entrada al modelo para el primer clúster.

A pesar de que las variables se comportan de forma similar, hay diferencias notorias en ambos clústeres. Para tener una mejor idea del contenido de los clústeres, se observan los promedios de cada una de las variables en ellos en la **Tabla 12**. Estos promedios, corresponden al centroide de cada clúster. El clúster 1 contiene los mayores valores de irradiancia y potencia. La temperatura de la celda y ambiente también son mayores. Con esta información se puede decir que el clúster 1 representa aquellos días donde la potencia es mayor, o días soleados, y el clúster 2 representa los días donde la potencia es más baja, o días nubosos.

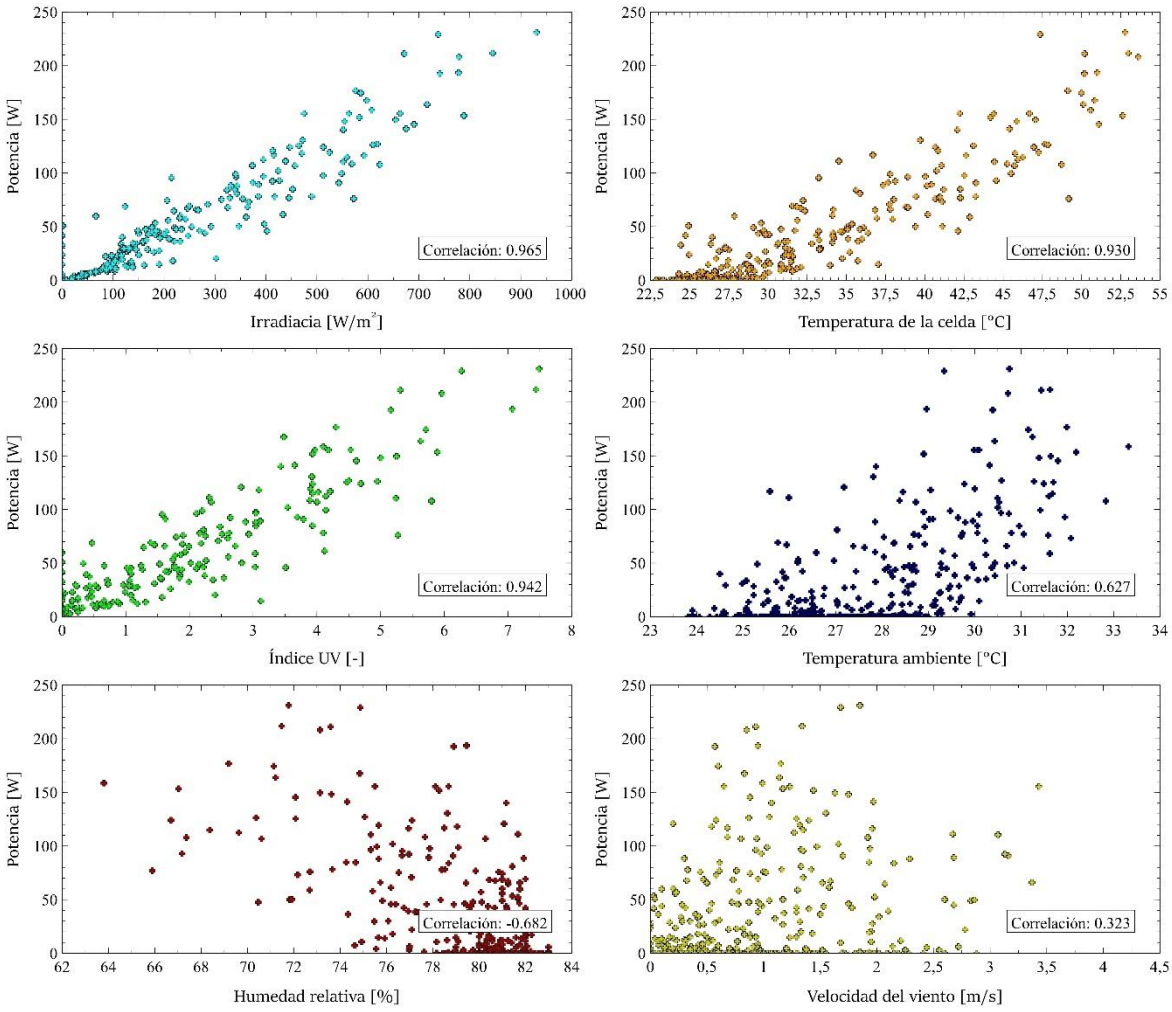

Figura 26 Matriz con gráficos de dispersión y correlaciones de la potencia contra las variables de entrada al modelo para el segundo clúster.

Tabla 12 Resumen de los valores de las variables dentro de cada clúster.

Variable	Clúster 1			Clúster 2		
	Mín.	Prom.	Máx.	Mín.	Prom.	Máx.
P [W]	0.000	47.51	247.61	0.000	27.31	231.04
G [W/m2]	0.000	197.98	966.06	0.000	121.56	931.20
UV	0.000	1.369	8.570	0.000	0.880	7.490
Tc [°C]	22.49	33.32	59.30	22.86	30.09	53.62
Ta [°C]	23.83	29.01	34.70	23.81	27.56	33.31
HR [%]	62.27	77.87	83.00	63.80	79.86	83.00
VV [m/s]	0.000	1.265	4.031	0.000	1.034	4.102

En la **Figura 27** se encuentra el día promedio para los valores de irradiancia y potencia de los tres grupos de datos. Estos grupos son: un grupo que contiene todos los datos que ya han sido tratados, un grupo que

contiene los datos del clúster 1, y otro grupo que contiene los datos del clúster 2. De ahora en adelante, las variables en estos grupos serán identificadas con los subíndices *T*, *1* y *2* respectivamente. Se puede ver con claridad que el trabajo del algoritmo k-means en este caso fue seleccionar del conjunto total de datos, C_T, aquellos con altos valores diarios de irradiancia, potencia y temperaturas, y bajos valores diarios de humedad relativa. Estos datos fueron etiquetados bajo la categoría de C_1. El resto de los datos, ubicados en C_2, representan valores mucho menores de irradiancia y potencia, como se ve con claridad en la **Figura 27**. En esta figura se encuentra el día promedio dentro de cada uno de los clústeres, y el día promedio para todos los datos.

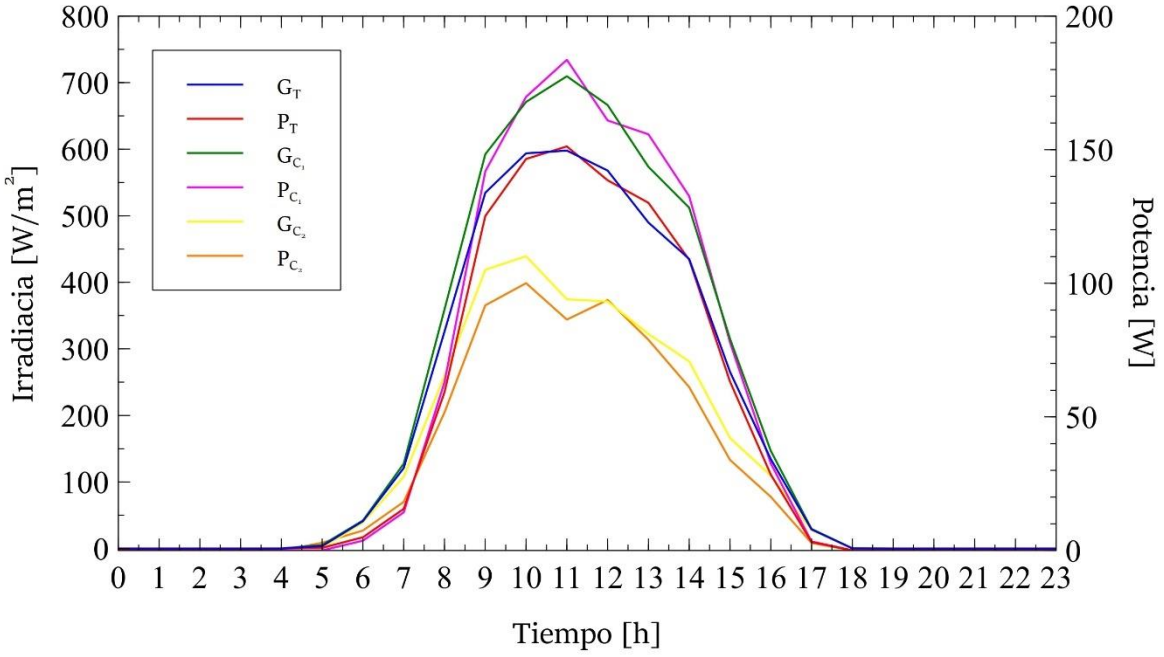

Figura 27 Día promedio de irradiancia y potencia para el grupo de datos completo (marcado como C_T), el clúster 1 (marcado como C1) y el clúster 2 (marcado como C2).

Capítulo 4: Descripción del modelo de predicción

Las redes neuronales son modelos de aprendizaje de máquinas que se inspiran en el sistema nervioso. En nuestro cerebro tenemos un gran número de neuronas que se encuentran altamente conectadas. Estas neuronas se dividen en tres partes, las dendritas, el cuerpo de la célula y el axon (Aggarwal, 2018). Las dendritas reciben señales eléctricas que llevan al cuerpo de la célula. Esta información es almacenada y compartida con otras neuronas por medio del axon. El punto de conexión entre una neurona y otra se llama sinopsis. Estas sinopsis evolucionan en el tiempo, haciéndose más débiles o fuertes dependiendo del uso que se les dé. A pesar de que las redes neuronales artificiales no se acercan a la complejidad que hay en el cerebro, hay similitudes entre las redes neuronales biológicas y las artificiales. En ambas redes hay componentes sencillos que forman la red y están altamente interconectados. Además, las conexiones entre una neurona y otra determinan la función que tiene una red (Hagan & Demuth, 2014).

4.1. Desarrollo del modelo matemático

Para empezar con la descripción matemática del modelo de redes neuronales de esta investigación se plantea el caso más simple de las redes neuronales. La forma más sencilla de una red neuronal tiene una sola neurona y también una sola entrada. Esta neurona se puede ver en la **Figura 28**.

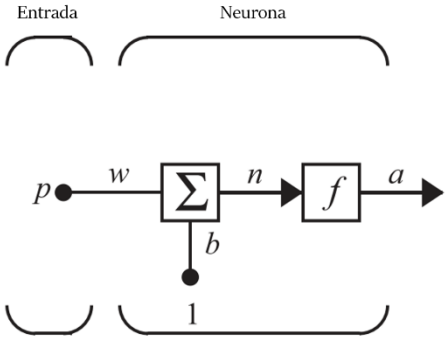

Figura 28 Neurona con una sola entrada. Adaptado de (Hagan & Demuth, 2014).

Donde p es la entrada de la red neuronal, w es el peso de la entrada, b es el término de sesgo, n es la suma ponderada de la neurona, f es la función de activación y a es la salida de la red neuronal. Esta salida puede ser escrita en los términos anteriores de la siguiente forma,

$$a = f(w \cdot p + b) \qquad \textbf{Ec. 16}$$

Existen varios tipos de funciones de activación lineales y no lineales. El trabajo de la función de activación en la mayoría de los casos es eliminar la linealidad de la suma ponderada que se lleva a cabo en la red neuronal, por ello, las funciones sigmoide y tangente hiperbólica son altamente utilizadas. La función de activación también cumple la función de hacer que la salida de la red neuronal se encuentre en un determinado rango de valores. Generalmente se busca que las funciones de activación tengan derivadas sencillas, puesto que esto reduce el costo computacional. Dependiendo de la tarea que una red neuronal deba cumplir, se deben elegir las funciones de activación de ésta. En la **Tabla 13** se pueden ver las funciones de activación más comunes.

Tabla 13 Funciones de activación de las redes neuronales más comunes. Adaptado de (Hagan & Demuth, 2014).

Nombre de la función	Relación entre la entrada y salida
Escalón	$a = 0 \quad n < 0$ $a = 1 \quad n \geq 0$
Escalón simétrico	$a = -1 \quad n < 0$ $a = 1 \quad n \geq 0$
Lineal	$a = n$
Lineal saturada o limitante	$a = 0 \quad n < 0$ $a = n \quad 0 \leq n \leq 1$ $a = 1 \quad n > 1$
Lineal positiva	$a = 0 \quad n < 0$ $a = n \quad n \geq 0$
Sigmoide	$a = \frac{1}{1 + e^{-n}}$
Tangente hiperbólica	$a = \frac{e^{n} - e^{-n}}{e^{n} + e^{-n}}$

Si en vez de tener una sola entrada, se tienen varias, el escalar p debe ser reemplazado por un vector $\mathbf{p}$ de longitud R, donde R es el número de entradas. De ahora en adelante, en las variables la notación será la siguiente: los escalares se muestran en minúscula o mayúscula sin negrita ni cursiva, los vectores en minúscula con negrita y sin cursiva, y las matrices en mayúscula con negrita y sin cursiva. Por otro lado, el peso en la **Ec. 16** dejará de ser un escalar también para ser una matriz, $\mathbf{W}$, que para el caso de una sola neurona será de una fila y R columnas. Esta ecuación en forma matricial se reescribiría como,

$$a(n) = f(\mathbf{W} \cdot \mathbf{p} + b) \qquad \textbf{Ec. 17}$$

Donde,

$$n = w_{1,1} \cdot p_1 + w_{1,2} \cdot p_2 + \cdots + w_{1,R} \cdot p_R + b \qquad \textbf{Ec. 18}$$

Ahora bien, generalmente una neurona no es suficiente para resolver un problema determinado. Las neuronas en una red se pueden conectar de varias formas. Pueden operar en paralelo, formando una capa. Pero, por ejemplo, una red neuronal que tenga una función de activación sigmoide en la primera capa y una función lineal en la segunda se puede entrenar para que se aproxime a casi cualquier función. Una sola capa de neuronas no puede hacer esto (Hagan & Demuth, 2014). Si se tienen S neuronas formando una capa, el vector de entradas debe estar conectado a cada una de las S neuronas. Esto da como resultado S matrices de pesos, cada una de dimensiones $1 \times R$. Pero por simplicidad se hablará de una sola matriz de pesos de dimensiones $S \times R$. El término de sesgo pasa de ser un escalar a ser un vector con longitud S. La suma ponderada de la neurona y la salida de la neurona también se convierten en vectores de longitud S. En la **Figura 29** se puede ver una red neuronal de una capa con S neuronas en ella y R entradas. La notación que se usa es abreviada, en vez de representarse cada neurona en un bloque, se muestran todas en uno solo y las dimensiones de la red se expresan en el tamaño de los parámetros. Los parámetros de una red neuronal se definen como los valores que se estiman en el proceso de entrenamiento. Estos son la matriz de pesos y el vector de sesgo. La red neuronal de la **Figura 29** es conocida como el *perceptrón*, esta red neuronal tuvo mucha popularidad desde su creación en 1958, sin embargo, no se tenían herramientas eficientes para poderlo entrenar. Fue en 1963 que se desarrolló el algoritmo de retropropagación de los errores que permitió volver al estudio de las redes neuronales. Una modificación de este algoritmo de aprendizaje es el que se utiliza en esta investigación, como se muestra más adelante.

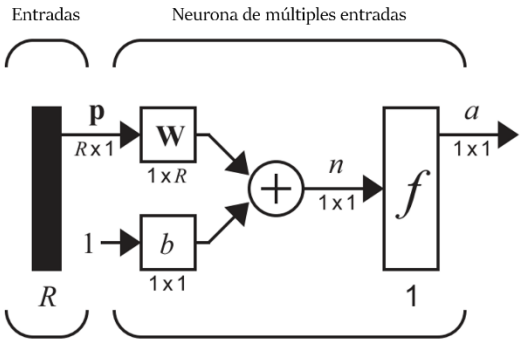

Figura 29 Red neuronal de una capa de S neuronas en notación abreviada. Adaptado de (Hagan & Demuth, 2014).

La **Ec. 17** entonces se reescribe como,

$$\mathbf{a(n) = f(W \cdot p + b)} \qquad \textbf{Ec. 19}$$

A esta estructura de conexión de las neuronas se le suele llamar red neuronal retroalimentada (*feedforward neural network*), ya que la información se mueve desde las entradas hasta la respuesta de la red, no hay bucles de información. En las redes neuronales de múltiples capas, se utiliza la misma conexión entre las entradas y la primera capa que se vio anteriormente. Es decir, la salida de una capa es la entrada de la siguiente capa. Otro nombre de esta estructura es *perceptrón multicapa*. En una red neuronal como en la de la **Figura 30**, hay tres capas de neuronas. Esta figura se encuentra en notación abreviada también, en la **Figura 5** está la notación extendida de esta red. A estas capas se les da el nombre de capas ocultas.

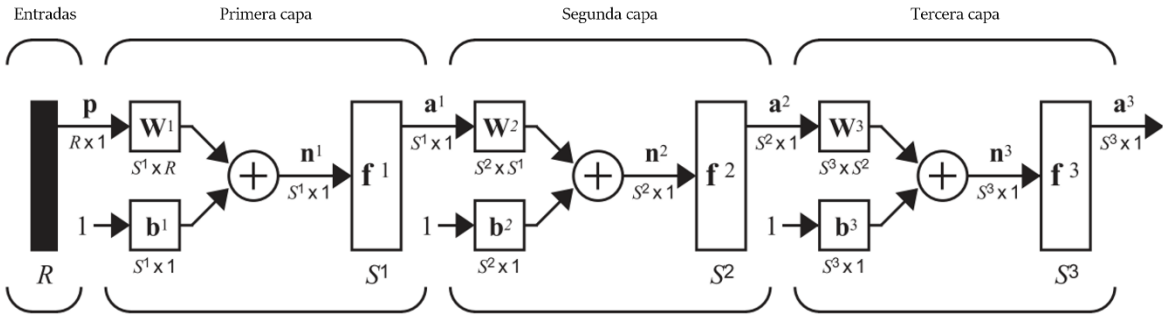

Figura 30 Red neuronal de tres capas con notación abreviada. Adaptado de (Hagan & Demuth, 2014).

Se puede ver que la primera la salida de una capa de neuronas es la entrada de la siguiente capa. Esto se ve expresado de forma genérica en la **Ec. 20**.

$$\mathbf{a}^{m+1} = \mathbf{f}^{m+1}(\mathbf{W}^{m+1} \cdot \mathbf{a}^{m} + \mathbf{b}^{m+1}) \qquad \textbf{Ec. 20}$$

La **Ec. 20** se cumple para $m = 0, 1, \dots, M - 1$. Donde M es el número total de capas. La primera capa oculta de neuronas recibe el número uno, y la capa de las entradas recibe el número cero. Para el caso particular de la primera capa de neuronas no se cumple la ecuación anterior. En cambio, se tiene,

$$\mathbf{a}^{0} = \mathbf{p} \qquad \textbf{Ec. 21}$$

La salida de las neuronas en la última capa se considera la respuesta de la red neuronal, y se define de la siguiente forma,

$$\mathbf{a} = \mathbf{a}^{M} \qquad \textbf{Ec. 22}$$

Para el caso de la **Figura 30** la salida de la primera capa se escribe como,

$$\mathbf{a}^{1} = \mathbf{f}^{1}(\mathbf{W}^{1} \cdot \mathbf{p} + \mathbf{b}^{1}) \qquad \textbf{Ec. 23}$$

En la segunda capa, la entrada de las neuronas es la **Ec. 23**. Al escribirlo en una ecuación se tiene lo siguiente,

$$\mathbf{a}^{2} = \mathbf{f}^{2}(\mathbf{W}^{2} \cdot \mathbf{a}^{1} + \mathbf{b}^{2}) \qquad \textbf{Ec. 24}$$

Finalmente, en la última capa, la salida de la red neuronal, la entrada de información es la salida de la segunda capa como se muestra en la **Ec. 25**.

$$\mathbf{a}^{3} = \mathbf{f}^{3}(\mathbf{W}^{3} \cdot \mathbf{a}^{2} + \mathbf{b}^{3}) \qquad \textbf{Ec. 25}$$

La **Ec. 25** puede escribirse en los términos de las **Ec. 23** y **Ec. 24** de la siguiente forma,

$$\mathbf{a}^{3} = \mathbf{f}^{3}(\mathbf{W}^{3} \cdot \mathbf{f}^{2}(\mathbf{W}^{2} \cdot \mathbf{f}^{1}(\mathbf{W}^{1} \cdot \mathbf{p} + \mathbf{b}^{1}) + \mathbf{b}^{2}) + \mathbf{b}^{3}) \qquad \textbf{Ec. 26}$$

Habiendo definido la estructura de la red neuronal, se explica a continuación el paso a paso y las ecuaciones necesarias para el proceso de aprendizaje de la red neuronal.

4.1.1. Algoritmo de retropropagación resiliente

En la red neuronal de la **Figura 30** se tienen como parámetros las tres matrices de pesos y los tres vectores de sesgo. La forma en la que se hayan estos parámetros es por medio de iteraciones en las que la red neuronal aprende de los datos que se le entregan. Existen varias técnicas de aprendizaje que permiten entrenar la red, el más utilizado por su robustez es el de propagación hacia atrás o retropropagación (*backpropagation*) (Ding, Wang, & Bi, 2011). En este algoritmo se entrena una red neuronal al asignar valores aleatorios para los parámetros, para luego ajustarlos en cada iteración. El algoritmo de propagación hacia atrás recibe su nombre porque propaga el error hacia atrás, es decir, de la última capa hasta la primera, ajustando los pesos con base en la responsabilidad que cada neurona tuvo sobre el error de la respuesta. Este método se puede describir como una generalización del algoritmo de mínimos cuadrados. La semejanza se encuentra en que ambos algoritmos actualizan pesos y ganancias con base en el error medio cuadrático. (Ponce Cruz, 2010). La popularidad de esta técnica se debe a que permite tener un método de optimización que se encuentra al definir el gradiente del error y minimizarlo con respecto a los parámetros de la red neural (Ponce Cruz, 2010).

Los datos que se le asignan a la red neuronal en el entrenamiento corresponden al concepto de aprendizaje supervisado. Es decir, sirven de instrucción a la red para saber, con un conjunto de datos de entrada, cuál es la salida que se espera que ella entregue. Esos datos ejemplo de cómo se debe comportar la red se pueden escribir como,

$$\{p_1, t_1\}, \{p_2, t_2\}, \dots, \{p_q, t_q\} \qquad \textbf{Ec. 27}$$

Donde p_q es una entrada de la red y t_q es su valor objetivo correspondiente. Para ajustar los pesos, el algoritmo compara la respuesta de la red neuronal con el valor objetivo. Esto lo hace con función del error, o función de coste. Generalmente, como función del error se utiliza el error cuadrático medio que se define en la siguiente ecuación de forma escalar.

$$E = e^2 = (t - a)^2 \qquad \textbf{Ec. 28}$$

La función de coste también puede ser escrita como una aproximación de la ecuación anterior en forma matricial para cada iteración k.

$$\widehat{E} = \big(\mathbf{t}(k) - \mathbf{a}(k)\big)^T \big(\mathbf{t}(k) - \mathbf{a}(k)\big) = \mathbf{e}^T(k)\mathbf{e}(k) \qquad \textbf{Ec. 29}$$

El objetivo del entrenamiento es lograr minimizar el error, al variar los parámetros de la red neuronal. Esto se hace con el algoritmo del descenso del gradiente (*gradient descent*). El algoritmo del descenso del gradiente utiliza la derivada de la función de coste con respecto a los parámetros de la red para actualizar el valor de los parámetros en cada iteración. Esta derivada es un vector conocido como el gradiente. La forma en la que se realiza la actualización de la matriz de pesos en forma escalar con el descenso del gradiente se encuentra en la siguiente ecuación.

$$w_{i,j}^m(k+1) = w_{i,j}^m(k) - \alpha \cdot \frac{\partial \widehat{E}}{\partial w_{i,j}^m} \qquad \textbf{Ec. 30}$$

Para el vector de sesgo se escribe en cambio,

$$b_i^m(k+1) = b_i^m(k) - \alpha \cdot \frac{\partial \widehat{E}}{\partial b_i^m} \qquad \textbf{Ec. 31}$$

Donde α es conocido como el ratio de aprendizaje. El valor de α tiene un impacto grande en el entrenamiento de la red neuronal, puesto que se puede interpretar como el ancho de los pasos que el algoritmo da para encontrar el mínimo de la función.

A pesar de su popularidad, este método tiene como desventaja la necesidad de asignar un valor para el ratio de aprendizaje. Para valores muy grandes el algoritmo no converge y para valores muy pequeños se requieren demasiadas iteraciones. Además, es un algoritmo que fácilmente cae en un mínimo local, tiene un convergencia generalmente lenta y es propenso a hacer muchas oscilaciones (Ding et al., 2011). El algoritmo de retropropagación resiliente (Rprop) en cambio, controla la actualización de los pesos para cada conexión durante el proceso de aprendizaje para minimizar las oscilaciones y maximizar el tamaño del paso de actualización (Igel & Hüsken, 2000).

En Rprop, a cada peso $w_{i,j}$ se le asigna un valor de actualización individual $\Delta_{i,j}$. Por simplicidad, el parámetro de sesgo se trata en las ecuaciones como si fuera el peso de una entrada extra. El término $\Delta_{i,j}$ se utiliza meramente para determinar el tamaño de la actualización de la matriz de pesos y cambia en el proceso de aprendizaje acorde a la siguiente regla (Riedmiller & Braun, 1993).

$$\Delta_{i,j}(k) = \begin{cases} \eta^+ \cdot \Delta_{i,j}(k-1), & \text{si } \frac{\partial E}{\partial w_{i,j}}(k-1) \cdot \frac{\partial E}{\partial w_{i,j}}(k) > 0 \\ \eta^- \cdot \Delta_{i,j}(k-1), & \text{si } \frac{\partial E}{\partial w_{i,j}}(k-1) \cdot \frac{\partial E}{\partial w_{i,j}}(k) < 0 \\ \Delta_{i,j}(k-1), & \text{en los demás casos} \end{cases} \qquad \textbf{Ec. 32}$$

El valor de $\Delta_{i,j}$ se encuentra limitado por dos barreras, que son Δ_{min} y $\Delta_{máx}$. Los condicionales en la **Ec. 32** indican el cambio en el signo de la derivada de la función de error con respecto a un cambio en el peso $w_{i,j}$. Si la derivada cambia de signo, indicando que el ajuste anterior fue muy grande y por ende el error se ha saltado un mínimo local, entonces $\Delta_{i,j}(k)$ decrece con el factor η^-. Si lo contrario ocurre, es decir, el signo de la derivada en la iteración anterior y en la actual son iguales, $\Delta_{i,j}(k)$ se aumenta con el factor η^+ para acelerar la convergencia del algoritmo (Riedmiller & Braun, 1993).

Después de haber adaptado el valor de $\Delta_{i,j}(k)$, se debe actualizar el valor de $w_{i,j}$ con las siguientes reglas. Si la derivada de la función de error con respecto a un cambio en el peso $w_{i,j}$ para la iteración k es positiva, es porque el error está aumentando, el peso se disminuye en función de $\Delta_{i,j}(k)$. Lo anterior se puede resumir en la siguiente ecuación.

$$\Delta w_{i,j}(k) = \begin{cases} -\Delta_{i,j}(k), & \text{si } \frac{\partial E}{\partial w_{i,j}}(k) > 0 \\ +\Delta_{i,j}(k), & \text{si } \frac{\partial E}{\partial w_{i,j}}(k) < 0 \\ 0, & \text{en los demás casos} \end{cases} \qquad \textbf{Ec. 33}$$

Ahora bien, el peso se actualiza al sumar el valor de la iteración anterior con el valor de la actualización del peso que acaba de ser calculada.

$$w_{i,j}(k+1) = w_{i,j}(k) + \Delta w_{i,j}(k) \qquad \textbf{Ec. 34}$$

La **Ec. 34** se cumple para todos los casos excepto cuando la derivada cambió de signo con respecto a la iteración anterior. En este caso el cambio de signo se da porque el algoritmo falló en entrar en un mínimo local y en cambio, lo ha saltado. En este caso, la actualización del peso se revierte con la siguiente ecuación.

$$\Delta w_{i,j}(k) = -\Delta w_{i,j}(k-1) \qquad \textbf{Ec. 35}$$

Se espera que en la siguiente iteración la derivada vuelva a cambiar de signo, y para evitar que el algoritmo repita el paso de la **Ec. 35**, se debe evitar que haya una actualización del valor de actualización. Esto se logra con la siguiente expresión.

$$\frac{\partial E}{\partial w_{i,j}}(k-1) = 0 \qquad \textbf{Ec. 36}$$

El procedimiento del algoritmo de retropropagación resiliente entonces inicia al generar valores aleatorios para todos los parámetros de la red neuronal multicapa y calcular la salida de la red (**Ec. 20**). Al ser valores aleatorios, entonces la salida de la red también será aleatoria, esto hará que al comparar la respuesta de la red con el valor objetivo por medio de la función de coste (**Ec. 29**) en la primera iteración, se consiga un error alto. Luego, acorde al signo de la derivada del error y las reglas en **Ec. 32** y **Ec. 33** se actualizan los pesos con la **Ec. 34**. Esto puede verse en el siguiente algoritmo.

Algoritmo 1 Aprendizaje de con el método de retropropagación resiliente (Rprop). Tomado de (Riedmiller & Braun, 1993).

{
    Para todos los pesos y términos de sesgo **{**
        Si $\frac{\partial E}{\partial w_{i,j}}(k-1) \cdot \frac{\partial E}{\partial w_{i,j}}(k) > 0$ entonces **{**
            $\Delta_{i,j}(k) = \min(\Delta_{i,j}(k-1) \cdot \eta^{+}, \Delta_{máx}$
            $\Delta w_{i,j}(k) = -\text{signo}\left(\frac{\partial E}{\partial w_{i,j}}(k)\right) \cdot \Delta_{i,j}(k)$
            $w_{i,j}(k+1) = w_{i,j}(k) + \Delta w_{i,j}(k)$
        }
        Sino, si $\frac{\partial E}{\partial w_{i,j}}(k-1) \cdot \frac{\partial E}{\partial w_{i,j}}(k) > 0$ entonces **{**
            $\Delta_{i,j}(k) = \text{máx}(\Delta_{i,j}(k-1) \cdot \eta^{-}, \Delta_{min}$
            $w_{i,j}(k+1) = w_{i,j}(k) - \Delta w_{i,j}(k-1)$
            $\frac{\partial E}{\partial w_{i,j}}(k-1) = 0$
        }
        Sino, si $\frac{\partial E}{\partial w_{i,j}}(k-1) \cdot \frac{\partial E}{\partial w_{i,j}}(k) = 0$ entonces **{**
            $\Delta w_{i,j}(k) = -\text{signo}\left(\frac{\partial E}{\partial w_{i,j}}(k)\right) \cdot \Delta_{i,j}(k)$

```
15 |   |   |   W_{i,j}(k + 1) = w_{i,j}(k) + Δw_{i,j}(k)
16 |   |   }
17 |   }
18 | }
```

Antes de entrenar las redes neuronales, los datos fueron escalados entre 0 y 1. Esto se hace con el fin de mejorar el rendimiento de la red neuronal. En la mayoría de los casos hay variables que por naturaleza tienen valores mucho mayores a los de otras variables. Por ejemplo, en el conjunto de datos que se estudia en esta investigación la irradiancia está en el orden de centenares y por otro lado el índice UV no alcanza la decena. En la **Ec. 37** se detalla la ecuación utilizada para escalar los datos antes de entrenar la red neuronal.

$$x_{escalado} = \tilde{x} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad \textbf{Ec. 37}$$

4.2. Entrenamiento y validación de las redes neuronales

Como se describió anteriormente, en el entrenamiento del modelo se deben entregar unos valores objetivo (**Ec. 27**) con los que se calcula el error y se hacen cambios en los parámetros del modelo hasta llegar a un mínimo local del error. A este conjunto de datos con los que se ajustan los parámetros se le conoce con el nombre de *datos de entrenamiento*. El error que se calcula con este conjunto de datos no es una medida real del desempeño del modelo de predicción. Cada conjunto de datos contiene cierta cantidad de ruido proveniente del proceso de medición. Como el modelo ajustó sus parámetros con los datos de entrenamiento, parte del ruido de estos datos los ha aprendido el modelo. A esto se le conoce bajo el término de *sobreajuste*. Por ello, si se quisiera validar el modelo con cualquier grupo de datos que tengan un ruido diferente, el error será mayor. Podría decirse con base a lo anterior que el error calculado con el conjunto de entrenamiento es una medida optimista del error real del modelo. Es una práctica común y necesaria entonces, dividir los datos en un conjunto de entrenamiento y otro que sirva para validar el rendimiento del modelo ante nuevos datos. Este segundo conjunto de datos son datos que el modelo no ha visto anteriormente, y se le conoce como *datos de validación*. Existen varias formas de dividir los datos en entrenamiento y validación, la más común de éstas es conocida como *regresión lineal*. En la regresión lineal se hace una partición, generalmente aleatoria de los datos, de forma tal que un porcentaje sea de entrenamiento y el restante de validación. En la mayoría de los modelos de aprendizaje de máquinas se toma un 80% de los datos para el entrenamiento y un 20% para validación. Esta división tiene como ventaja lo sencilla y fácil de implementar que es, en parte a eso se debe su popularidad. No obstante, como desventaja se tiene que mientras más datos se elijan para entrenar el modelo, menos se tendrán para la validación, teniendo como resultado un modelo con un error muy bajo pero una baja validez también.

4.2.1. Validación cruzada de K iteraciones

Una solución a esta desventaja se encuentra en los métodos de validación cruzada, donde todos los datos se utilizan en entrenamiento y validación. Estos métodos son especialmente útiles cuando se cuenta con un número limitado de observaciones (James, Witten, Hastie, & Tibshirani, 2013). Entre ellos está el método *K-Fold*, en el que se dividen los datos en K conjuntos y el entrenamiento del modelo se realiza K veces. En

la primera iteración, se utiliza la unión de los datos dejando por fuera el conjunto K, para entrenar el modelo, y seguidamente se usa únicamente el conjunto K para la validación. Esto se repite hasta haber utilizado los K conjuntos en la validación del modelo. Al final, el método entrega los valores promedio de la medida de error o exactitud que se esté utilizando para todos los K conjuntos de validación, como se muestra en la **Ec. 38**. De esta manera se elimina, o al menos se reduce, el efecto de aleatoriedad en la elección del conjunto de entrenamiento concreto (Kubat, 2017), ya que se le da oportunidad a todos los datos para participar en el entrenamiento y validación. Cabe aclarar que este método es más robusto, pero requiere mayor poder computacional, puesto que se entrena el modelo K veces en vez de una como en el método de regresión lineal.

$$\mathrm{CV_K} = \frac{1}{\mathrm{K}} \cdot \sum_{\mathrm{i=1}}^{\mathrm{K}} \mathrm{RMSE_K} \qquad \textbf{Ec. 38}$$

Donde RMSE es la raíz del error cuadrático medio para el conjunto de validación en cada iteración K y $\mathrm{CV_K}$ es el RMSE promedio de los K conjuntos de validación. En la **Figura 31** se encuentra una representación gráfica de la validación cruzada para el caso K=5. El nombre del método adapta el valor de K, así que en ese caso se llamaría validación cruzada 5-Fold.

Figura 31 Representación gráfica de la partición de datos 5-Fold para validación cruzada.

En general, a K se le dan valores de 5 o de 10, pero no hay una regla formal que indique el valor más adecuado. A medida que K aumenta, la diferencia de tamaño entre el conjunto de entrenamiento y los subconjuntos de validación disminuye. Y a medida que esta diferencia disminuye, el sesgo de la técnica se vuelve más pequeño (Kassambara, 2017; Kuhn & Johnson, 2013). En este trabajo se utilizó K=10 para dividir el conjunto de datos (validación cruzada 10-Fold). En este enfoque, las divisiones de los K conjuntos de aproximadamente el mismo tamaño se hacen de forma aleatoria (James et al., 2013). Se ha demostrado que en modelos de predicción de potencia fotovoltaica es una buena práctica darle un orden aleatorio a los datos antes de hacer particiones (Dolara, Grimaccia, Leva, Mussetta, & Ogliari, 2018). Adicional al RMSE, se

estudian el nRMSE que se encuentra en la **Ec. 5** y el R^2 que se detalla a continuación. El R^2 es una proporción de la varianza total en la variable dependiente que se explica por las variables independientes (Keith, 2019).

$$R^2 = \frac{\sum(y_{pred} - \bar{y}_{pred}) \cdot (y_{med} - \bar{y}_{med})}{\sqrt{\sum(y_{pred} - \bar{y}_{pred})^2 \cdot \sum(y_{med} - \bar{y}_{med})^2}} \qquad \textbf{Ec. 39}$$

4.2.2. Ajuste de hiperparámetros

El proceso de entrenamiento se hizo ajustando los hiperparámetros de la red. Estos son el número de capas, número de neuronas en cada capa, la función de activación y los valores de η^- y η^+ para el algoritmo de aprendizaje Rprop. Como función de activación se utilizó la sigmoide en las capas ocultas y la función lineal en la salida de la red neuronal. Por otro lado, los valores de η^- y η^+ fueron 0.5 y 1.2 ya que son los valores recomendados por los creadores del algoritmo Rprop y siguen siendo utilizados exitosamente en estudios recientes (Riedmiller & Braun, 1993). Se sabe que rara vez se necesitan más de dos capas en la red neuronal para hacer regresiones de modelos complicados (Zhang, Chen, Malik, & Hope, 1993). Por ello el número de capas ocultas se fijó en dos, no obstante, el número de neuronas por capa fue una variable a estimar dentro del proceso de optimización de la red neuronal. En el siguiente algoritmo se muestran los pasos de esta optimización de los hiperparámetros.

Algoritmo 2 Ajuste de los hiperparámetros del modelo.

```
{
    Definir conjuntos de cada hiperparámetro del modelo para evaluar.
    Para cada conjunto de hiperparámetros {
        Para cada iteración haga {
            Extraer datos específicos.
            Entrenar el modelo en los datos restantes.
            Predecir con los datos extraídos.
        }
        Calcular el rendimiento promedio a través de las predicciones con los datos extraídos.
    }
    Determinar el conjunto de hiperparámetros óptimo.
    Ajustar el modelo a todos los datos de entrenamiento usando el conjunto de hiperparámetros
    óptimo.
}
```

Para la capa oculta 1, se establecieron 5, 4 y 3 neuronas. Y para la segunda capa oculta se establecieron 3, 2 y 1. Luego de haber implementado el **Algoritmo 2**, se comparan los errores que arroja cada conjunto de hiperparámetros a variar. El error más bajo es el que permite determinar el conjunto de hiperparámetros óptimo. En la **Figura 32** a la izquierda se encuentra el error promedio de validación para distintas combinaciones de número de neuronas con el C_1. En el eje horizontal está el número de neuronas de la capa 1. En cambio, las líneas indican los distintos números de neuronas de la capa 2.

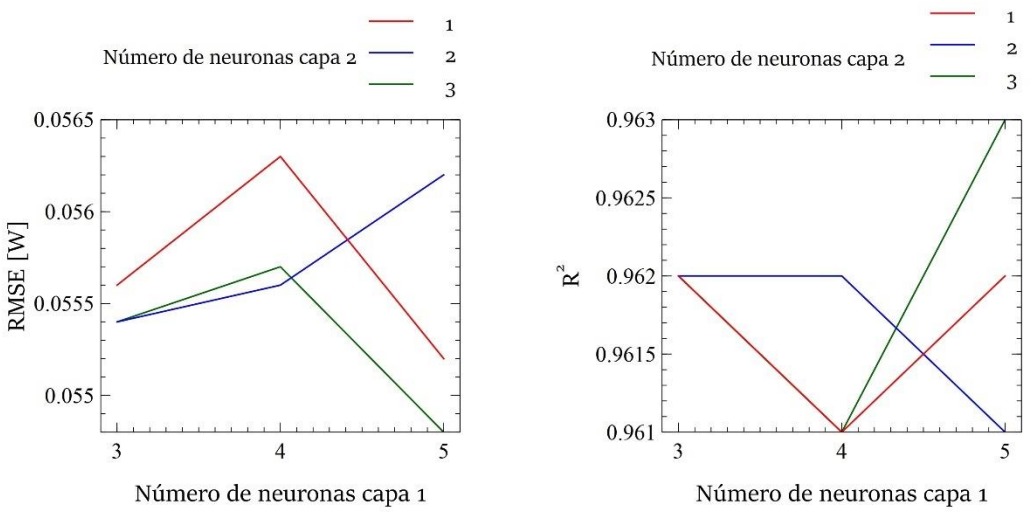

Figura 32 RMSE en la izquierda y R^2 en la derecha para distintos números de neuronas por capa en la primera y segunda capa ocultas de la red neuronal del C_1.

En esta red neuronal, el mínimo error se encuentra para 5 neuronas en la primera capa y 3 en la segunda. Cabe resaltar que los errores para las demás combinaciones de número de neuronas no son mucho más grandes. Por otro lado, en la **Figura 32** a la derecha se encuentra el R^2 promedio de validación para distintas combinaciones de número de neuronas con el C_1. Nuevamente, el óptimo se da para 5 neuronas en la primera capa y 3 en la segunda.

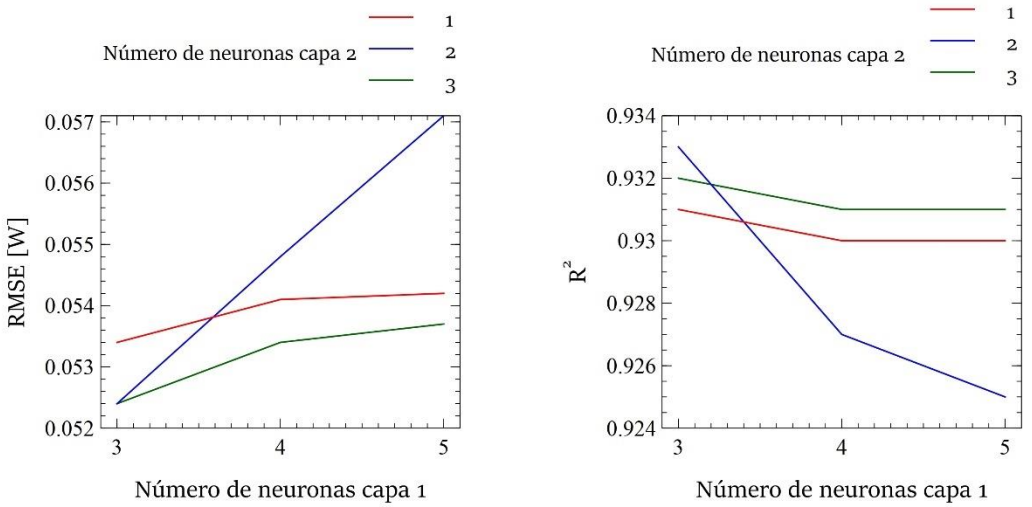

Figura 33 RMSE en la izquierda y R^2 en la derecha para distintos números de neuronas por capa en la primera y segunda capa ocultas de la red neuronal del C_2.

En la **Figura 33** y **Figura 34** se encuentran el RMSE y R^2 para distintos números de neuronas en la red neuronal del C_2 y del C_T. el mínimo valor de RMSE se encuentra para dos combinaciones de neuronas en cada capa. Estas son, para 3 neuronas en la primera capa y 2 en la segunda, y para 3 neuronas en la primera capa y 3 en la segunda. No obstante, el mayor valor de R^2 se obtiene es para la primera de estas combinaciones.

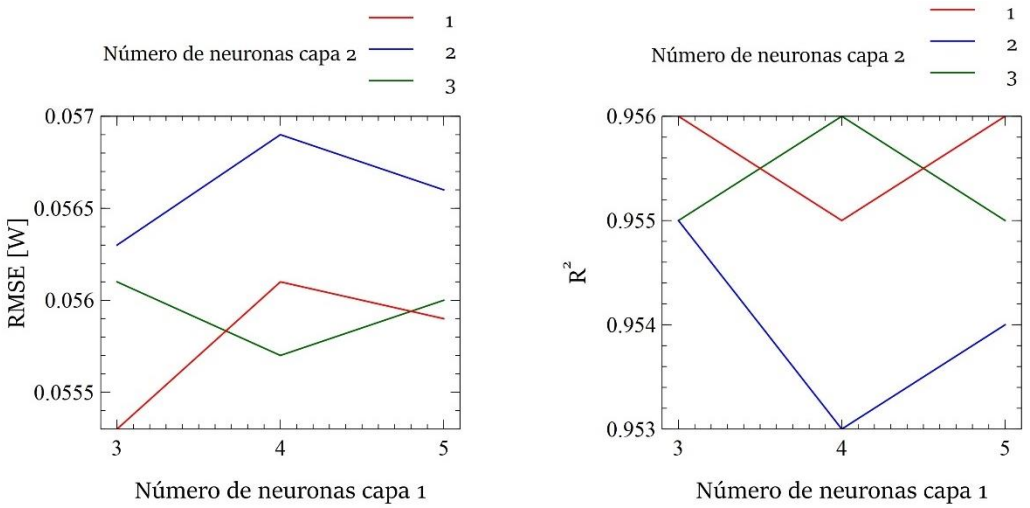

Figura 34 RMSE en la izquierda y R^2 en la derecha para distintos números de neuronas por capa en la primera y segunda capa ocultas de la red neuronal del C_T.

Por último, para C_T se obtiene un mínimo RMSE al utilizar tres neuronas en la primera capa y una en la segunda. En la siguiente tabla se encuentra un resumen de los hiperparámetros utilizados finalmente para entrenar los modelos de predicción. Como se puede ver, no se incluyó la velocidad del viento dentro de las variables para predicción. Esto se debe a que al incluirla se presentaban mayores errores. Lo mismo ocurrió para los tres modelos.

Tabla 14 Entradas e hiperparámetros de las redes neuronales entrenadas.

Nombre del modelo	Datos de la red neuronal	Entradas	Neuronas capa oculta 1	Neuronas capa oculta 2	Función de activación
NNC_T	C_T	G, Tc, Ta, UV y HR	3	1	Sigmoide/Lineal
NNC_1	C_1	G, Tc, Ta, UV y HR	5	3	Sigmoide/Lineal
NNC_2	C_2	G, Tc, Ta, UV y HR	3	2	Sigmoide/Lineal

En la **Figura 35** se muestra un esquema de cómo se conectan las neuronas en la red para el C_1.

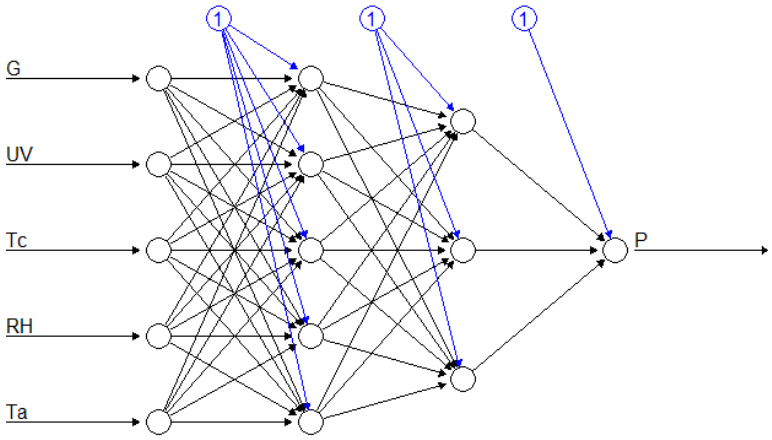

Figura 35 Representación gráfica de la red neuronal para C_1.

Por otro lado, al **Figura 36** representa la red neuronal para el C_2 y la **Figura 37** para el C_T.

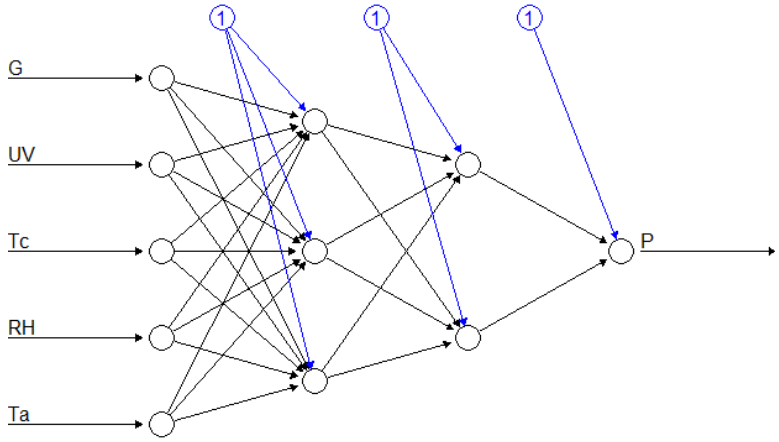

Figura 36 Representación gráfica de la red neuronal para el C_2.

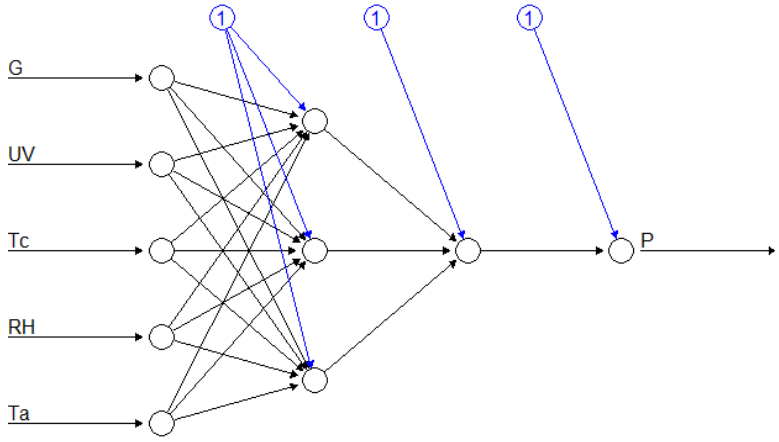

Figura 37 Representación gráfica de la red neuronal para C_T.

Un resumen de los valores obtenidos de nRMSE y R2 se encuentran en la **Tabla 15**. Cabe resaltar que, aunque los errores son similares en magnitud, se consiguió una reducción del error total al haber dividido los datos en dos grupos homogéneos. El menor nRMSE se tuvo en el clúster 2, siendo este 5.24%. En cambio, el mayor ajuste de R^2 se obtuvo en el clúster 1, donde se encuentran los días soleados.

Tabla 15 Errores de validación para los modelos de predicción con redes neuronales.

Modelo	nRMSE	R^2
NNC_1	5.48%	0.963
NNC_2	5.24%	0.933
NNC_T	5.53%	0.956

Antes de comparar gráficamente los resultados obtenidos en el modelo de predicción con los datos actuales, estos fueron llevados a su escala natural. En la siguiente ecuación se muestra el cálculo para llevar los datos

a su escala natural. Como las predicciones de potencia no tienen un valor máximo ni mínimo en escala natural, se han utilizado los valores del conjunto original de datos.

$$x = \tilde{x} \cdot (x_{max} - x_{min}) + x_{min} \qquad \textbf{Ec. 40}$$

En la **Figura 38** se encuentran los valores actuales del clúster 1, y los valores predichos con el modelo NNC_1. Se puede ver un alto ajuste como lo demuestran el nRMSE y el R^2. Sin embargo, en la mayoría de los picos de potencia, la predicción no alcanza al valor actual. En los demás puntos, los valores son tan cercanos que es difícil distinguir las líneas de predichos y actuales.

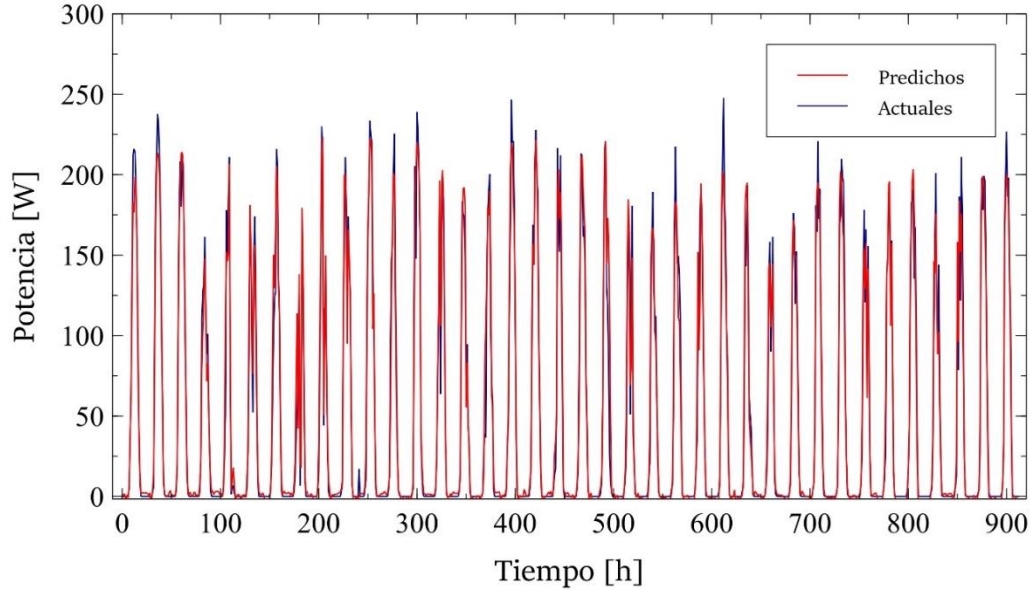

Figura 38 Potencia predicha y actual de los datos del C_1.

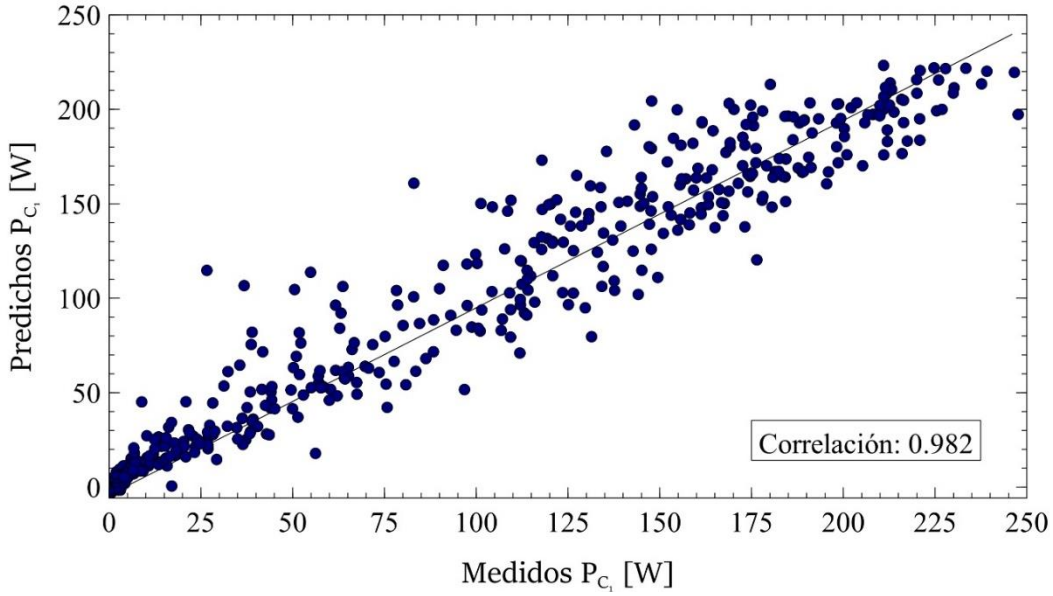

Figura 39 Gráfico de dispersión y correlación entre valores predichos y actuales para el C_1.

Por otro lado, en la **Figura 39** se encuentra la correlación de predichos versus actuales, que es de 0.982. La mayoría de los puntos se encuentran formando una línea diagonal de 45° indicando un buen ajuste del modelo de predicción. Para el caso del clúster 2, la potencia predicha y actual se muestran en la **Figura 40**. Nuevamente se puede ver un alto ajuste como lo demuestran el nRMSE y el R^2. Para este modelo los picos de potencia predichos son mucho más cercanos a la potencia actual en el clúster. Esta puede ser la razón por la que el modelo presenta un menor nRMSE.

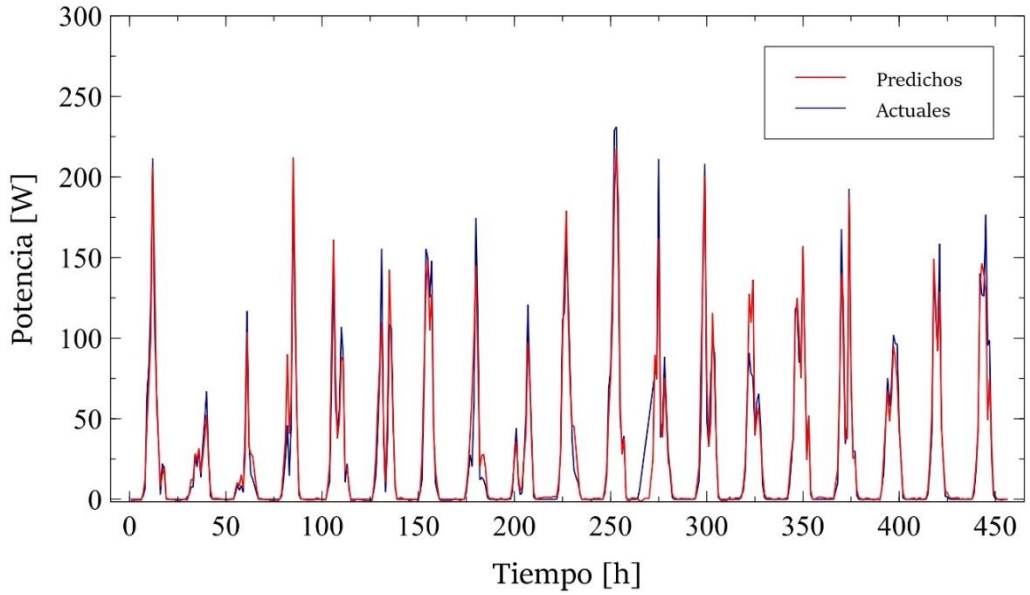

Figura 40 Potencia predicha y actual de los datos del C_2.

En la **Figura 41** se muestra la correlación entre valores predichos de potencia y los actuales para el clúster 2. Para este caso también hay una correlación alta entre estas dos variables.

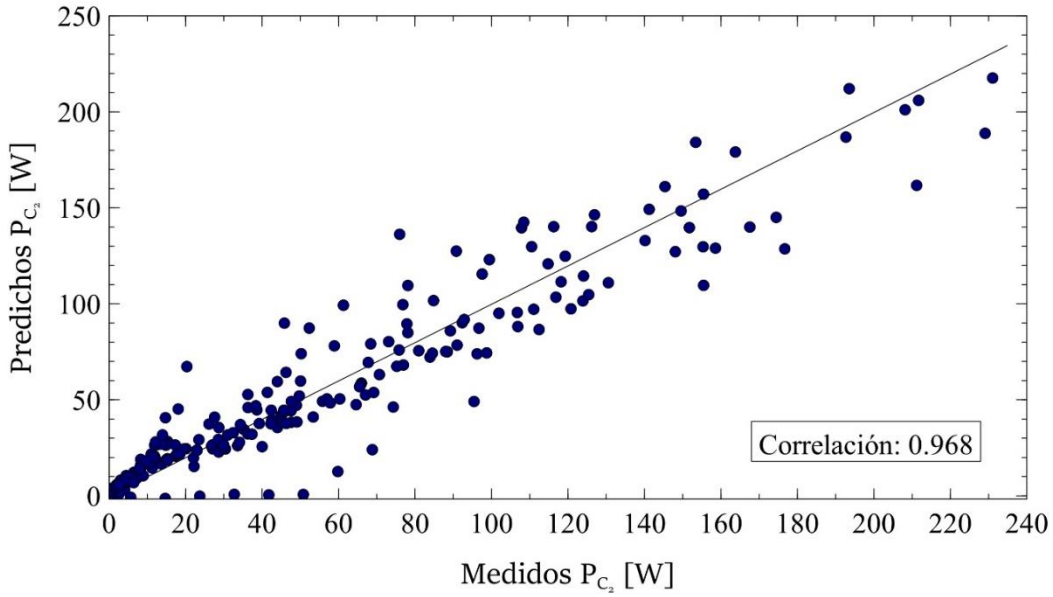

Figura 41 Gráfico de dispersión y correlación entre valores predichos y actuales para el C_2.

Finalmente, en la **Figura 42** se encuentran los valores predichos de potencia y los actuales para el conjunto total de datos. En este modelo ocurre los mismo que en el primero, los valores predichos no alcanzan a los actuales en el pico de potencia. Aun así, el error es pequeño dado que, para los demás datos, la cercanía es alta. En la **Figura 43** se encuentra el gráfico de dispersión con la correlación entre predichos y supuestos, que nuevamente es bastante alta.

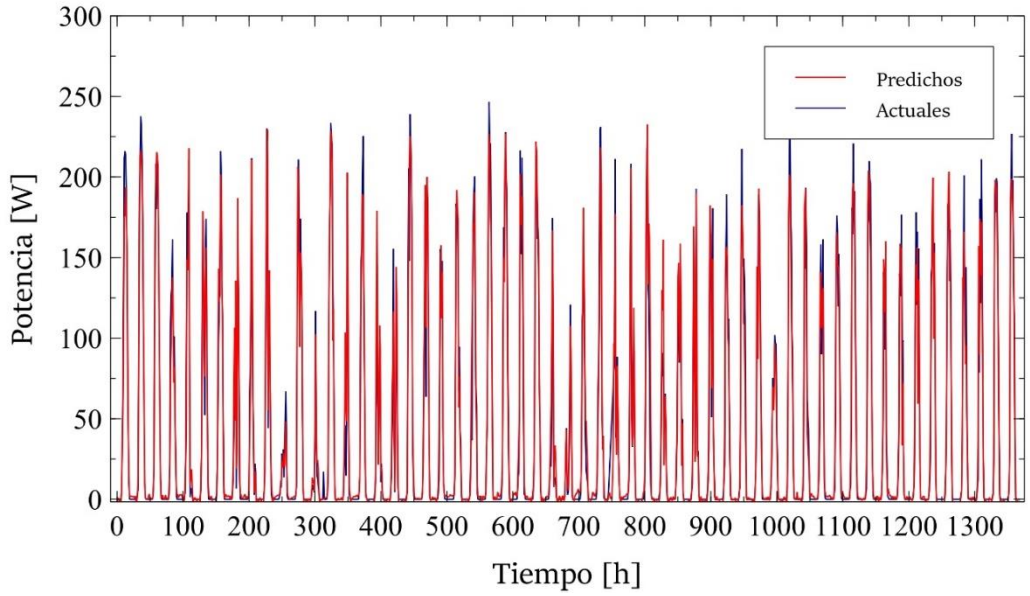

Figura 42 Potencia predicha y actual de los datos del C_T.

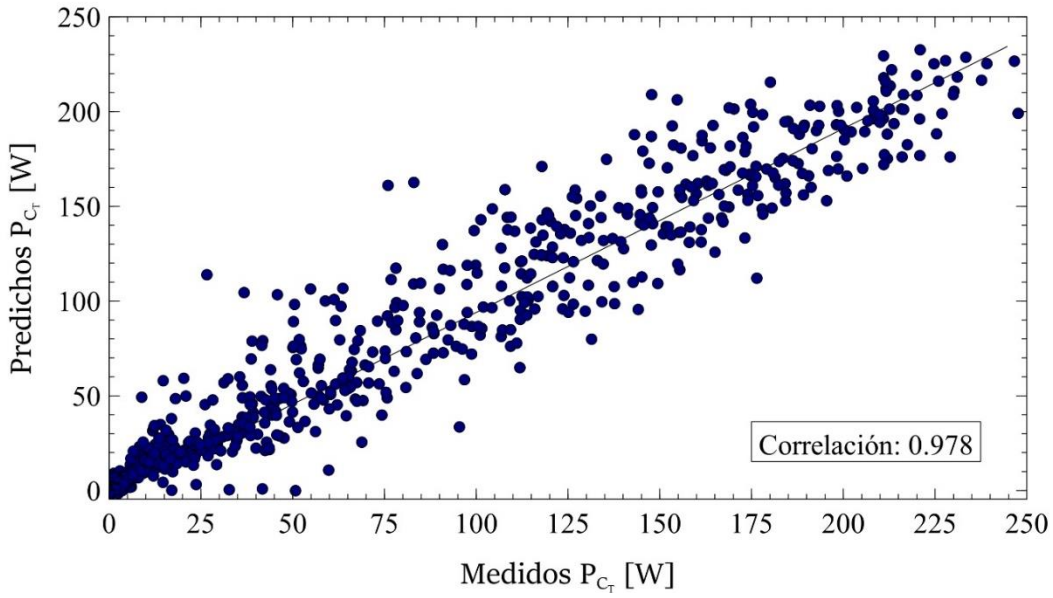

Figura 43 Gráfico de dispersión y correlación entre valores predichos y actuales para el C_T.

En la **Figura 44** se encuentra la superficie de respuesta para la potencia diaria en función de la irradiancia y la temperatura del conjunto total de datos. Los valores corresponden cercanamente a los datos medidos del conjunto total de datos.

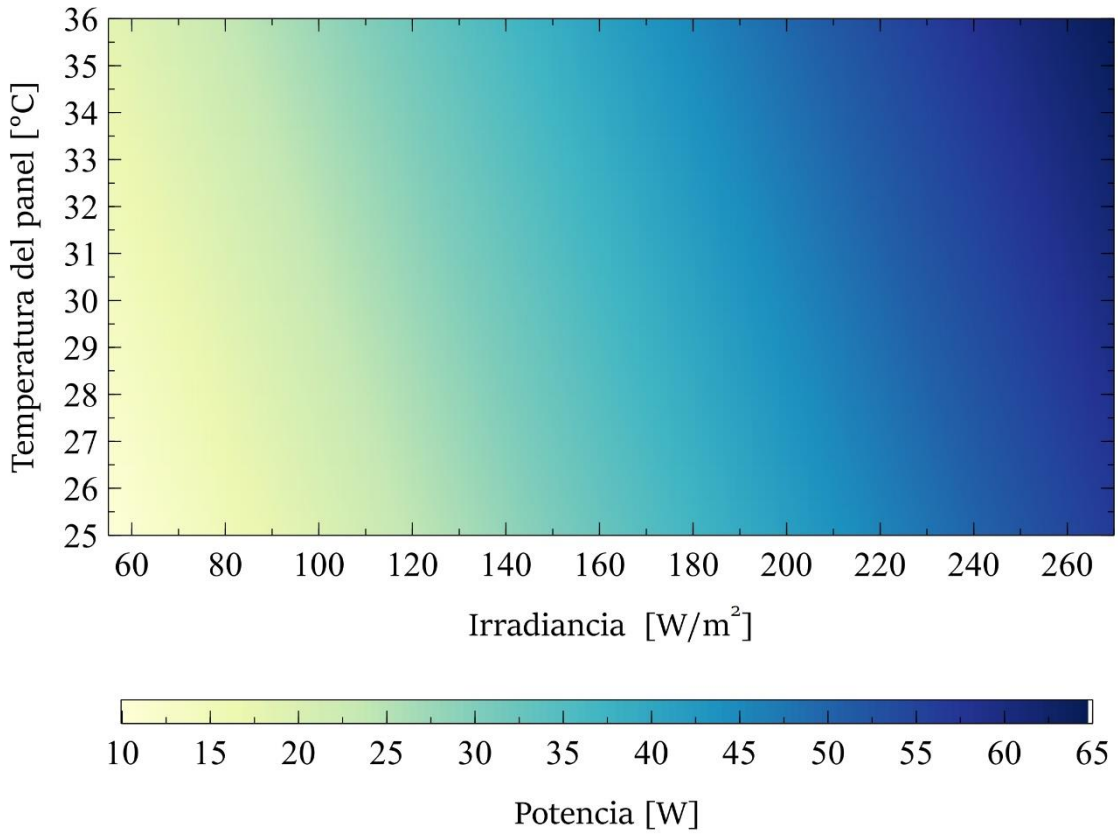

Figura 44 Gráfico de superficie de respuesta para la potencia diaria en función de la irradiancia y la temperatura con C_T.

Con base a la teoría que se ha explicado en este capítulo, es posible escribir en una ecuación los modelos que se han desarrollado. Esta ecuación es similar a la **Ec. 26**, teniendo como diferencia que en el algoritmo Rprop, se utiliza la matriz de pesos para llevar el vector de sesgo de cada capa. La ecuación de estos modelos es,

$$\mathbf{a}^3 = \mathbf{f}^3\left(\mathbf{W}^3 \cdot \mathbf{f}^2\left(\mathbf{W}^2 \cdot \mathbf{f}^1(\mathbf{W}^1 \cdot \mathbf{p})\right)\right) \qquad \textbf{Ec. 41}$$

Donde las funciones de activación $\mathbf{f}^2$ y $\mathbf{f}^1$ corresponden a la función sigmoide, y $\mathbf{f}^3$ correcponde a la función lineal. La definición de estas dos funciones se encuentra en la **Tabla 13**.

Los pesos $\mathbf{W}^1$, $\mathbf{W}^2$ y $\mathbf{W}^3$ se encuentran en a **Tabla 16** para cada uno de los conjuntos de datos. Cabe resaltar que tanto en la **Ec. 41** como en la **Tabla 16** todos los pesos incluyen el término de sesgo en la última fila de la matriz. Esto tiene como consecuencia que en el término $\mathbf{p}$ hay una última fila adicional con el término uno (1). De esta forma, al multiplicarse por la matriz de los pesos, se obtiene la misma respuesta que se obtendría en la **Ec. 26**.

Tabla 16 Pesos de los modelos NNC_1, NNC_2 y NNC_T. Dentro de los pesos se encuentra el vector de sesgo.

Modelo	$\mathbf{W^1}$	$\mathbf{W^2}$	$\mathbf{W^3}$
NNC_1	$\begin{bmatrix} -1.92 & 0.08 & 2.38 & 0.02 & -1.19 \\ 3.58 & 1.57 & 1.65 & 1.19 & 0.26 \\ 0.54 & -0.54 & 1.00 & -1.00 & 0.93 \\ 0.55 & 1.82 & -11.01 & -0.68 & 0.14 \\ -0.36 & -0.86 & -0.84 & -2.43 & -1.09 \\ -0.77 & -1.48 & 0.16 & -0.75 & 1.32 \end{bmatrix}$	$\begin{bmatrix} -1.18 & 0.88 & -0.08 \\ 2.50 & -1.05 & 2.69 \\ 0.73 & -3.07 & 1.06 \\ 0.43 & -1.12 & -0.32 \\ -3.60 & 1.29 & -3.70 \\ 1.29 & -1.08 & -0.10 \end{bmatrix}$	$\begin{bmatrix} -0.15 \\ 0.93 \\ -1.37 \\ 0.32 \end{bmatrix}$
NNC_2	$\begin{bmatrix} -6.18 & -2.09 & -0.18 \\ 3.42 & -0.75 & -1.20 \\ 3.89 & 2.61 & 6.14 \\ 0.87 & 0.53 & -0.15 \\ 0.31 & 0.10 & 0.16 \\ -2.92 & 1.07 & -0.43 \end{bmatrix}$	$\begin{bmatrix} -0.63 & -0.47 \\ -4.35 & 5.01 \\ -3.03 & 1.43 \\ -1.16 & 0.79 \end{bmatrix}$	$\begin{bmatrix} -0.98 \\ -0.29 \\ 1.96 \end{bmatrix}$
NNC_T	$\begin{bmatrix} 0.98 & 2.11 & 1.90 \\ 3.72 & 0.16 & -3.31 \\ -0.32 & -0.36 & -1.72 \\ -0.08 & -2.50 & 2.65 \\ -0.71 & -1.36 & 1.09 \\ -1.13 & 0.72 & -0.99 \end{bmatrix}$	$\begin{bmatrix} 1.02 \\ 1.60 \\ -2.27 \\ -1.73 \end{bmatrix}$	$\begin{bmatrix} -0.37 \\ 1.76 \end{bmatrix}$

Capítulo5: Rendimiento del modelo de predicción

En el primer capítulo de este documento se habló de la importancia de predecir potencia fotovoltaica y los beneficios que tiene sobre la producción de energía solar. Dependiendo del horizonte de predicción, el valor predicho puede ayudar a distintas planificaciones dentro de una planta fotovoltaica. En esta investigación, se desarrolló un modelo de predicción de potencia con redes neuronales optimizadas con el algoritmo Rprop. Este modelo no tiene dependencia del tiempo directamente. Para hallar un valor de potencia en el futuro, es necesario suministrar al modelo las variables independientes en ese tiempo del futuro. Una limitante fuerte de esta investigación fue la cantidad de datos disponibles para el entrenamiento y validación del modelo. Esta limitante va más allá del correcto entrenamiento del modelo, por el contrario, se encuentra en su capacidad de predicción. Es recomendable cuando se entrena un modelo de predicción, usar todos los datos que el modelo va a recibir en el futuro (Hagan & Demuth, 2014). Sin embargo, esto es muy difícil de garantizar. En la práctica común se recomienda al entrenar modelos de predicción de potencia, utilizar al menos los datos de un año. De esta forma se garantiza que el modelo ha aprendido de todas las estaciones y cuenta con días variados. En este caso se cuenta con poco menos de dos meses de medición, entre septiembre y noviembre del 2019. Esto hace que el horizonte de predicción no pueda ser por mucho tiempo al futuro. Además, obliga al modelo a predecir potencia solo entre los meses con los cuales fue entrenado. Con base en el estudio bibliográfico, y teniendo en cuenta las anteriores consideraciones, los modelos de predicción desarrollados en esta investigación para el clúster 1 y 2 tienen un horizonte de predicción dependiente del día en que se quiera predecir. Las predicciones tendrán un error menor siempre que el día a predecir sea similar a los utilizados en el entrenamiento de los modelos.

A pesar de la desventaja por tener un número reducido de datos, como resultado del tratamiento la agrupación de estos, se tienen errores bajos en la validación de los modelos ante datos nuevos. En la siguiente tabla se encuentran los errores me predicción de potencia conseguidos por diferentes investigadores en años recientes.

Tabla 17 Errores de predicción encontrados en otros documentos de la literatura científica.

Ref.	Horizonte de predicción	Error de predicción	Modelo	Descripción
(Haque, Nehrir, & Mandal, 2013)	1 día	MAPE 3.38%-11.83%, nRMSE 12.11%-13.13%	WT, Fuzzy ARTMAP	Modelo de predicción híbrido en el que inicialmente se filtran los datos con WT, se

				clasifican con y se desarrolla el modelo de predicción.
(De Giorgi et al., 2014)	1 h a 24 h	nRMSE 10.91%-23.99%	MR y Elman ANN	En este modelo MR se usó para hacer un análisis de sensibilidad y Elman ANN para predecir potencia.
(Leva, Dolara, Grimaccia, Mussetta, & Ogliari, 2017)	24 h	nRMSE 12.5%-36.9%	ANN	Se usó el modelo de día claro y se predijo potencia para días soleados, nublado y parcialmente nublado. También se hizo un análisis de sensibilidad de la red neuronal. La exactitud del modelo está altamente relacionada con el procesamiento de los datos.
(Zhu et al., 2016)	1 día	nRMSE 7.193%-19.663%	WD y ANN	Se filtran las perturbaciones en la potencia utilizando WD y luego se predice potencia utilizando redes neuronales artificiales.
(Cheng et al., 2017)	-	MAPE 20.7%-34.5%	BP-NN	En este modelo se dividen los datos en varios clústeres antes de utilizar redes neuronales basadas en back-propagation para predecir. Encontraron que el modelo funciona mejor cuando los datos han sido agrupados.
(Z. Liu et al., 2020)	600 min	nRMSE 2.84%-7.19%	EML-Chicken swarm optimizer	Se clasifican los datos en día soleado, lluvioso, o nublado. Luego, se realizan modelos de predicción con redes neuronales para cada categoría optimizando con el método de enjambres de gallinas llegando a errores de predicción pequeños para
(Ramsami & Oree, 2015)	24 h	MAE 2.09%-2.31%	GR y ANN	En el modelo se utilizaron en conjunto la regresión generalizada y las redes neuronales para predecir potencia. La regresión se utilizó para determinar las variables relevantes y su importancia.
(Pedro & Coimbra, 2012)	2 h	nRMSE 13.97%-18.71%	ANN optimizadas con GA	En este documento se compara el desempeño de las redes neuronales optimizadas con GA, con modelos ARIMA, kNN y ANN. Se comprueba que las redes neuronales tienen un error inferior, y que los algoritmos genéticos son de ayuda en la optimización de la red.

Descripción de las siglas: WT: Wavelet transform, Fuzzy ARTMAP: Fuzzy logic and adaptive resonance theory. MR: Multiple regression. ANN: Artificial neural network. WD: wavelet decomposition. NP-NN: Backpropagation based neural network. ELM: Extreme learning machine. GR: Generalized regression. GA: Genetic algorithm.

Por consecuente, la metodología desarrollada en esta investigación presenta una ventaja en la creación de modelos que permitan predecir potencia. En este trabajo no solo se comparan los errores de predicción de los modelos para los clústeres con el modelo que contiene todos los datos. Además de esto, se revisan los

modelos más comunes presentes en la literatura, que son el de persistencia y el de regresión lineal multivariada. En las siguientes secciones se detalla cada uno de estos.

5.1. Modelo de persistencia

El modelo de persistencia es utilizado mayormente en la literatura como un modelo de referencia (Rana & Rahman, 2020; Rodríguez-Benítez et al., 2020). En este modelo, se asume que la potencia en una hora, t, es igual a la potencia a esa misma hora en el día anterior. En otras palabras, el modelo de persistencia solo se basa en la correlación lineal entre los valores de energía fotovoltaica presentes y los futuros (Colak et al., 2020). Es evidente que este error tiene errores grandes puesto que el clima no es constante. No obstante, es una buena aproximación en casos donde no se ha ajustado ningún modelo o no se cuenta con datos para hacerlo. Para el conjunto total de datos, la distancia media cuadrática mínima normalizada encontrada para el modelo de persistencia fue de 17.81%. La gráfica de este modelo se puede ver a continuación, es notorio que cada día es una copia del anterior, por lo que en días donde la irradiancia fue baja, pero el día anterior alta, se tienen altas potencias.

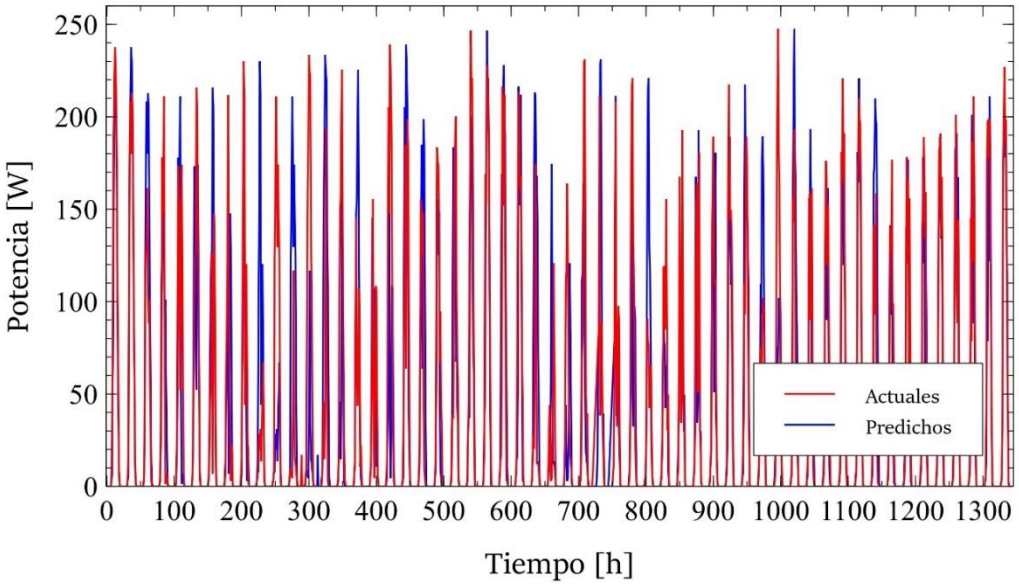

Figura 45 Potencia predicha y actual en el modelo de persistencia.

5.2. Regresión lineal múltiple

La regresión multivariada es una buena herramienta para entender la relación entre variables y para desarrollar modelos de predicción. En este caso se intenta escribir la potencia fotovoltaica en función de las variables que se han estado estudiando a lo largo de esta investigación. Para encontrar los parámetros de este modelo hace una partición por regresión lineal, como se explicó en el capítulo anterior. Una regla general en la partición de datos es tomar el 80% de los datos para entrenar el modelo, y el 20% restante para hacer la validación. Una buena práctica es dividir los datos en entrenamiento y validación de forma tal

que estos dos conjuntos sean similares. Así se garantiza también que, al menos de los datos que se tienen para entrenar la red, no habrá datos extremos que no se utilizarán en el entrenamiento. En la **Figura 46** se muestra la distribución de los datos para entrenamiento y validación y se puede ver cómo ambos grupos son similares en las densidades de datos que tienen. Ambos conjuntos poseen la mayor cantidad de datos para valores bajos de potencia. También, ambos grupos tienen dentro de sí el valor máximo de potencia. Para entrenar esta red se destinó el 80% de los datos a entrenamiento y el 20% restante a validación. Esta partición dejó como resultado 1094 datos para el entrenamiento y 274 para la validación.

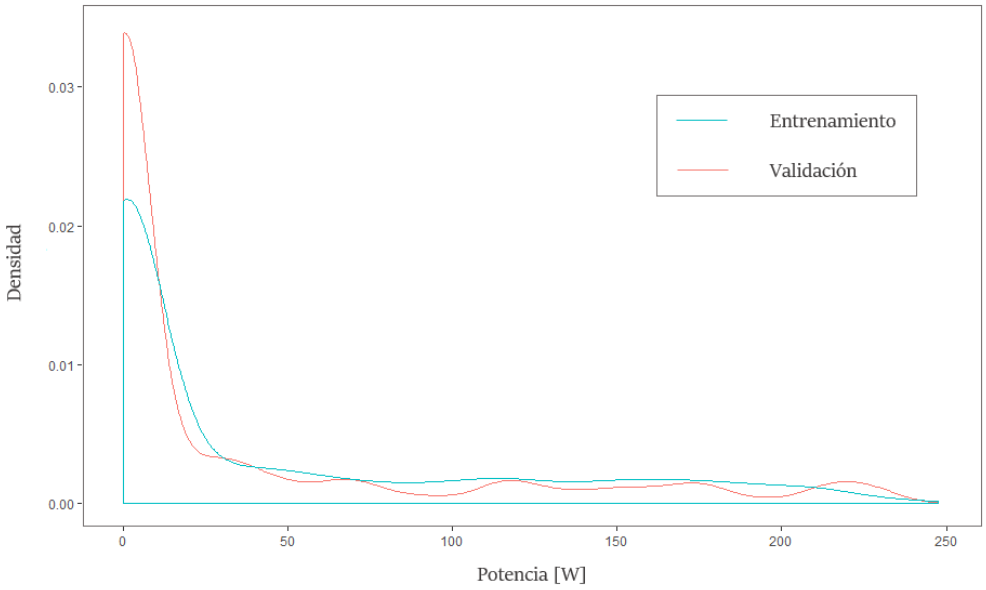

Figura 46 Distribución de los datos de entrenamiento y validación para C_T.

De igual forma, en la **Tabla 18** se muestra el promedio, la desviación estándar y el estadístico z de los datos que se encuentran en cada grupo para el conjunto de datos completo. El estadístico z corresponde a la media aritmética del conjunto de datos dividida por la desviación estándar. Para ambos grupos, los valores de los datos son similares como se podía observar en la figura anterior.

Tabla 18 Resumen de los datos en el grupo de entrenamiento y validación para C_T.

Variable	Promedio de entrenamiento	Promedio de validación	Desviación estándar entrenamiento	Desviación estándar validación	Estadístico z entrenamiento	Estadístico z validación
P [W]	42.4	34.4	62.5	62.8	0.650	0.549
G [W/m²]	178	149	260	257	0.687	0.580
UV	1.24	1.06	1.99	2.03	0.624	0.525
Tc [°C]	32.5	31.2	8.80	8.38	3.69	3.73
Ta [°C]	28.6	28.3	2.16	2.12	13.2	13.4
HR [%]	78.4	78.9	4.40	4.17	17.8	18.9
VV [m/s]	1.11	1.10	0.752	0.794	1.48	1.38

La obtención de los parámetros del modelo de regresión lineal se hizo con el método de mínimos cuadrados. Se comprobó la significancia de cada uno de los coeficientes en el modelo, siendo el intercepto significativo con un valor de -85.6. Para valores de irradiancia iguales a cero, se espera que la potencia también sea cero. Sin embargo, se tienen valores negativos cuando la irradiancia es cero. De igual forma, se encontró que los coeficientes de temperatura ambiente, velocidad del viento y humedad relativa no son significativos. Los coeficientes de este modelo se hallaron después de hacer escalado los datos con la **Ec. 37**. Los coeficientes se encuentran en la siguiente tabla ANOVA.

Tabla 19 Análisis de varianza para los coeficientes del modelo de regresión lineal múltiple.

Variable	Coeficiente	Std. Error	Valor t	Pr(>|t|)
Intercepto	-85.6	5.28	-16.21	<2e-16 ***
G	0.138	0.007	20.71	<2e-16 ***
Tc	3.182	0.197	16.15	<2e-16 ***

Códigos de significancia: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

El nRMSE de este modelo fue de 5.98% y el modelo de regresión simultánea resultante tiene la siguiente forma.

$$P = -85.6 + 0.138 \cdot G + 3.182 \cdot T_c \qquad \textbf{Ec. 42}$$

En la **Figura 47** se encuentra la gráfica con los valores predichos y actuales del modelo de regresión multivariada. La potencia predicha toma valores negativos cuando se espera que sea cero. De igual forma, en muchos de los picos, la potencia predicha no alcanza a la actual. En la mayoría de los valores, sobre todo cuando la potencia se encuentra aumentando o disminuyendo, los valores son tan cercanos que no se alcanza a ver diferencia entre la línea de predichos y la de valores actuales.

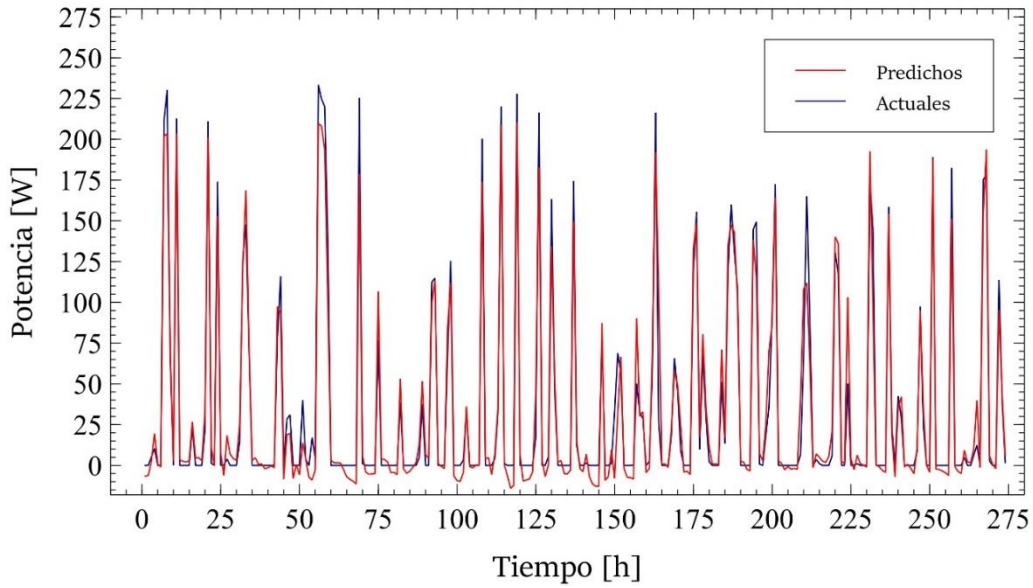

Figura 47 Potencia predicha y actual de los datos de validación del modelo de regresión lineal multivariada.

Ahora bien, verificar la significancia de los coeficientes del modelo no es suficiente para garantizar que un modelo de regresión lineal es adecuado. Además, se deben comprobar los supuestos de normalidad de los residuos, homocedasticidad e independencia. En la **Figura 48** se pueden ver cuatro gráficas que ayudan a comprobar los supuestos. Es claro que la normalidad no se cumple (gráfico arriba a la derecha) y que esto no se debe a valores atípicos, invalidando los resultados obtenidos. Para solucionar este problema se puede hacer una transformación de la variable de respuesta. Sin embargo, son pocas las que se pueden aplicar dado que la variable de respuesta del modelo contiene ceros.

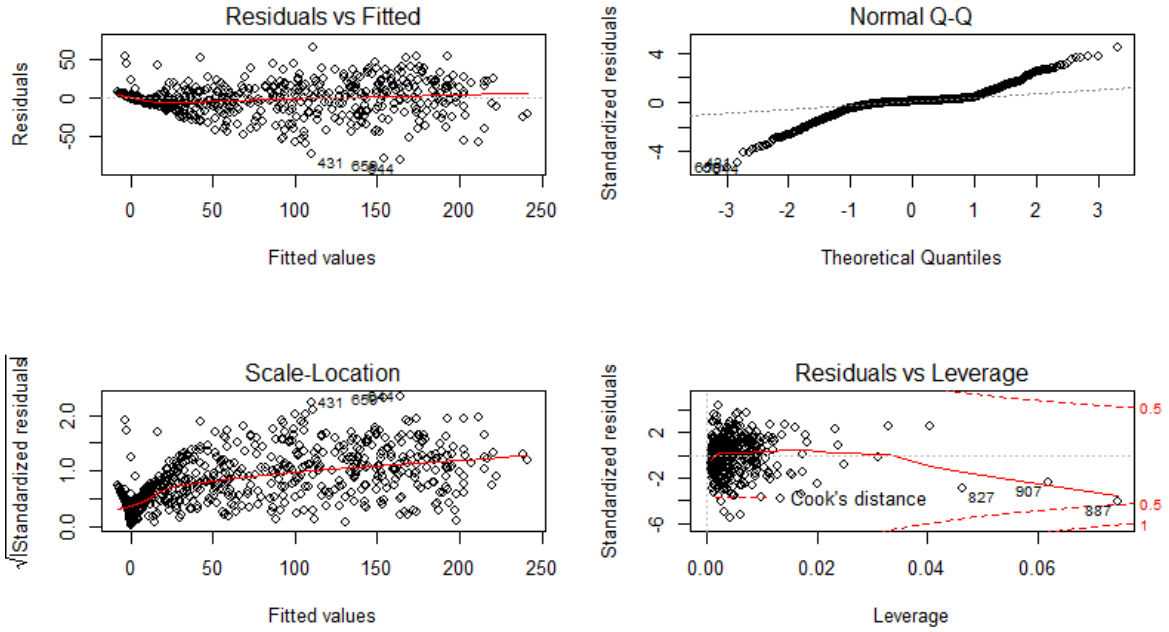

Figura 48 Verificación de los supuestos del modelo de regresión lineal múltiple.

La regresión lineal tiene como desventaja que su estructura está sujeta a supuestos que no siempre se pueden cumplir. A pesar de que las regresiones lineales con un caso específico de las redes neuronales, estas últimas pueden aproximar una gran cantidad de modelos sin que se tengan que hacer hipótesis entre las variables dependientes e independientes. Por el contrario, en las redes, estas relaciones se determinan es en el proceso de aprendizaje del modelo.

Finalmente, lo anterior deja como resultado el modelo de persistencia únicamente como base de referencia para comparar los modelos de redes neuronales desarrollados en el **Capítulo 4**. Al comparar los 4 errores se tiene que el de persistencia por bastante el mayor, con un 17.81%. Esto era de esperarse puesto que es un modelo poco robusto que asume que el clima entre un día y el siguiente no varía. Sin embargo, es de utilidad como modelo de referencia para verificar si un modelo más complejo, es útil o no. Seguidamente, se encuentra la red neuronal multicapa para entrenada con C_T que obtiene un error de validación de 5.53%. Por último, los modelos de redes neuronales tienen un desempeño mayor, siendo los errores para los clústeres 1 y 2 5.24% y 5.48%, respectivamente, menores al error de la red entrenada con el conjunto total de datos.

Capítulo 6: Conclusiones y trabajos futuros

6.1. Conclusiones

La predicción de potencia juega un papel fundamental en las plantas de generación fotovoltaica. Se desea cada vez tener predicciones más confiables, y con un mayor horizonte de predicción. En este documento se desarrolló una metodología de ajuste de modelos de predicción, que dio como resultado tres modelos de predicción de potencia fotovoltaica a corto plazo con redes neuronales, y un modelo de persistencia. El desarrollo de los modelos fue en base horaria, con datos extraídos de una plataforma experimental ubicada en la Universidad del Norte. El total de datos disponibles fue examinado a fin de elegir solo los datos que representaran información relevante al modelo de predicción de potencia.

El tratamiento de los datos tiene un papel fundamental en los resultados del modelo de predicción. Es necesario siempre antes de entrenar un modelo, hacer un análisis exhaustivo de los datos que se tienen y cuáles de estos son solo perturbaciones que aumentarán el error de validación de este. Por ello, se eligieron cuidadosamente varios filtros que se aplicaron sobre las observaciones. Como los modelos a entrenar fueron redes neuronales, este tratamiento tiene en especial más sentido, ya que al tener datos con mucho ruido se podría presentar el caso de sobreajuste en el que la red deja de generalizar y su capacidad de predicción es pobre ante nuevos datos. El resultado de estos filtros en los datos fue obtener un error bastante menor a los que se encuentran en la literatura.

La imputación de datos fue una herramienta fundamental para no descartar observaciones con información relevante para el modelo. Se pudo comprobar que la imputación por medio de interpolación lineal es una técnica sencilla que para pocos datos no afecta la media de las variables.

En este documento se eligieron las variables independientes por medio de análisis de correlación. Se pudo ver que variables como la velocidad del viento, con correlaciones entre 0.314 y 0.334 entre ellas y la potencia fotovoltaica, no son de aporte al modelo de predicción. En cambio, incluir variables con una alta correlación, como la irradiancia (con correlaciones en los distintos conjuntos de datos de 0.965) y la temperatura de la celda (con correlaciones en los distintos conjuntos de datos de entre 0.930 y 0.958), sí ayuda a desarrollar un modelo que reproduce bien los datos.

Por otro lado, los parámetros encontrados y la configuración de los modelos de redes neuronales son válidos en Puerto Colombia, Colombia. También, estos resultados son válidos únicamente para los días pertenecientes a los meses en los cuales se hicieron las mediciones de los datos utilizados en el entrenamiento de estos modelos de predicción. Para días soleados la red neuronal que tuvo el menor error

constó de 5 neuronas en la primera capa oculta y 3 en la segunda. Esto dio como resultado un nRMSE de validación promedio del 5.48%, R^2 de 0.963 y un coeficiente de correlación entre valores predichos y actuales de 0.982. Para los días nubosos, la configuración óptima de la red fue 3 neuronas en la primera capa oculta y 2 en la segunda. Entregando como resultado un nRMSE de validación promedio de 5.24%, un R^2 de 0.933 y un coeficiente de correlación del modelo de 0.968. Estos valores de nRMSE se encuentran muy por debajo de los valores que se encuentran en la literatura. Para el conjunto con todos los datos el error fue del 5.53%, el R^2 de 0.965 y el coeficiente de correlación entre valores predichos y actuales de 0.978. Esto se consiguió con una red más sencilla que las anteriores, teniendo 3 neuronas en la primera capa oculta y una en la segunda.

Por lo anterior, se concluye que agrupar los datos en días soleados y nubosos ayuda a reducir el error de predicción del modelo. Para el caso estudiado en esta investigación se contó con poco menos de dos meses con días pertenecientes a los meses de septiembre a noviembre. Al no tener datos de todo un año, no se tiene una variedad alta de días en el conjunto de datos disponible. Esto hace que el conjunto global de datos tenga baja variabilidad. Y a pesar de conseguir dos clústeres diferentes entre sí, los datos en ellos no son tan lejanos como serían en caso de tener datos para un año completo. A esto se le atribuye la similitud en los nRMSE de los modelos que se encontró.

La metodología desarrollada mostró que se puede reducir el error de predicción de potencia fotovoltaica. sin embargo, se requieren más datos para tener un mejor entrenamiento de las redes neuronales y poder predecir un día cualquiera del año. Ahora bien, el horizonte de predicción fijado en este documento corresponde a una limitada cantidad de datos para entrenar los modelos. Aunque se desea que las predicciones sean de días o semanas, esto solo es posible teniendo gran cantidad de datos que le permitan al modelo estar lo suficientemente bien entrenado. Al momento de entrenar una red neuronal, es necesario asegurarse de que los datos que se utilizan en el entrenamiento sean un reflejo de todos los datos con los que el modelo se va a encontrar posteriormente. Esto es muy difícil de garantizar, sin embargo, se recomienda que los datos pertenezcan a varias estaciones del año, y sean medidos de forma rigurosa para disminuir el ruido. Es decir, deben representar bien el fenómeno físico que se desea predecir.

Cabe aclarar que el tamaño del set de validación no está relacionado con el horizonte de predicción del modelo. Lo único que permite expandir el tiempo a futuro al que se hacen las predicciones es el contar con un modelo más robusto, es decir, que haya sido entrenado con gran variedad de datos. O bien, varios modelos para cada tipo de día. Al tener pocos datos, se validó la efectividad del método de validación cruzada 10-Fold. Esta herramienta permitió no incurrir en el *trade-off* de asignar un porcentaje al tamaño de datos para entrenamiento y validación. En caso de contar con muchos datos, la validación 10-Fold puede volverse un proceso que requiere gran poder computacional. Por ello, se recomienda en ese caso utilizar la partición de los datos implementada en este documento para el modelo de regresión lineal múltiple.

Consecuentemente, se concluye que las redes neuronales son una herramienta que permite recrear una gran variedad de funciones, independientemente de la relación entre las variables. Sin embargo, dificultan la comprensión del fenómeno que se estudia a través del modelo. Esto se debe a que no existe un vínculo simple entre los pesos y la función que se aproxima. Por eso se les da el nombre de caja negra. A pesar de que su implementación matemática no es compleja en muchos casos, es difícil hallar relaciones entre

variables a través de este tipo de modelos. En los casos donde se quieren estudiar estas relaciones entre variables independientes y dependientes, es necesario valerse de más herramientas estadísticas para entender las relaciones entre variables. Para hacer únicamente predicción de potencia, las redes neuronales han demostrado ser una herramienta que se desempeña de forma excelente.

Finalmente, el algoritmo de Rprop cumplió con el objetivo de garantizar la rápida convergencia en el entrenamiento de los modelos. El no tener un tamaño de paso fijo (ratio de aprendizaje) en la búsqueda del error mínimo ayudó a entrenar los modelos mucho más rápido que con el algoritmo de retropropagación basado en el descenso del gradiente. De igual forma, limitar las redes neuronales a geometrías sencillas permitió evitar el problema de sobreajuste de los modelos.

6.2. Trabajos futuros

Dentro de los trabajos futuros se tiene primeramente poder probar con una cantidad mucho mayor de datos la hipótesis de que la agrupación de estos lleva a disminuciones en el error de predicción. De igual forma, al tener más días de mediciones, se puede estudiar la forma en la que se agrupan los días de distintas estaciones en Puerto Colombia, Colombia.

También, se recomienda estudiar el efecto de combinar distintas funciones de activación en las distintas capas de la red neuronal, y utilizar más capas en estas. Esto con el fin de hacer un análisis más amplio del ajuste de los hiperparámetros. De igual forma, se recomienda estudiar el efecto del uso de distintas metodologías de filtrado y agrupación de los datos en el resultado final de la red neuronal.

Asimismo, las redes neuronales recurrentes juegan un papel importante en esta tarea de predicción. Permitir que la red neuronal tenga memoria ayuda en la tarea del estudio de las series de tiempo. Se recomienda como trabajo futuro aplicar la misma metodología de tratamiento y agrupación de datos, pero utilizando redes como LSTM o GRU para la predicción de potencia con distintos horizontes de predicción.

Finalmente, como trabajo futuro se recomienda el desarrollo de un modelo de predicción de temperatura del panel solar y contrastarlo con los que se encuentran en la literatura.

Bibliografía

Abuella, M., & Chowdhury, B. (2015). Solar power probabilistic forecasting by using multiple linear regression analysis. *SoutheastCon 2015*, *2015-June*(June), 1–5. https://doi.org/10.1109/SECON.2015.7132869

Aggarwal, C. C. (2018). Neural Networks and Deep Learning. En *Artificial Intelligence*. https://doi.org/10.1007/978-3-319-94463-0

Amral, N., Ozveren, C. S., & King, D. (2007). Short term load forecasting using Multiple Linear Regression. *2007 42nd International Universities Power Engineering Conference*, *2018-Janua*, 1192–1198. https://doi.org/10.1109/UPEC.2007.4469121

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Boehmke, B. (2019). K-means Cluster Analysis. Recuperado el 16 de febrero de 2020, de University of Cincinnaty website: https://uc-r.github.io/kmeans_clustering

Bouzerdoum, M., Mellit, A., & Massi Pavan, A. (2013). A hybrid model (SARIMA-SVM) for short-term power forecasting of a small-scale grid-connected photovoltaic plant. *Solar Energy*, *98*(PC), 226–235. https://doi.org/10.1016/j.solener.2013.10.002

Cambridge Spark. (2019). Tutorial: Introduction to Missing Data Imputation. Recuperado el 18 de noviembre de 2020, de Medium website: https://medium.com/@Cambridge_Spark/tutorial-introduction-to-missing-data-imputation-4912b51c34eb

Cheng, K., Guo, L. M., Wang, Y. K., & Zafar, M. T. (2017). Application of clustering analysis in the prediction of photovoltaic power generation based on neural network. *IOP Conference Series: Earth and Environmental Science*, *93*(1). https://doi.org/10.1088/1755-1315/93/1/012024

Cleveland, C., & Morris, C. (2014). Handbook of Energy VOLUME II: Chronologies, Top Ten Lists, And

Word Clouds. En *Elsevier*. Elsevier.

Colak, M., Yesilbudak, M., & Bayindir, R. (2020). *Daily Photovoltaic Power Prediction Enhanced by Hybrid GWO - MLP , ALO - MLP and WOA - MLP Models Using Meteorological Information*. 1–18. https://doi.org/10.3390/en13040901

da Silva Fonseca, J. G., Oozeki, T., Takashima, T., Koshimizu, G., Uchida, Y., & Ogimoto, K. (2012). Use of support vector regression and numerically predicted cloudiness to forecast power output of a photovoltaic power plant in Kitakyushu, Japan. *Progress in Photovoltaics: Research and Applications*, *20*(7), 874–882. https://doi.org/10.1002/pip.1152

Das, U. K., Tey, K. S., Seyedmahmoudian, M., Mekhilef, S., Idris, M. Y. I., Van Deventer, W., … Stojcevski, A. (2018). Forecasting of photovoltaic power generation and model optimization: A review. *Renewable and Sustainable Energy Reviews*, *81*(August 2017), 912–928. https://doi.org/10.1016/j.rser.2017.08.017

De Giorgi, M. G., Congedo, P. M., & Malvoni, M. (2014). Photovoltaic power forecasting using statistical methods: Impact of weather data. *IET Science, Measurement and Technology*, *8*(3), 90–97. https://doi.org/10.1049/iet-smt.2013.0135

Diagne, M., David, M., Lauret, P., Boland, J., & Schmutz, N. (2013). Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews*, *27*, 65–76. https://doi.org/10.1016/j.rser.2013.06.042

Ding, M., Wang, L., & Bi, R. (2011). An ANN-based Approach for Forecasting the Power Output of Photovoltaic System. *Procedia Environmental Sciences*, *11*(3), 1308–1315. https://doi.org/10.1016/j.proenv.2011.12.196

Dinov, I. D. (2018). Data science and predictive analytics: Biomedical and health applications using R. En *Data Science and Predictive Analytics: Biomedical and Health Applications using R*. https://doi.org/10.1007/978-3-319-72347-1

Dolara, A., Grimaccia, F., Leva, S., Mussetta, M., & Ogliari, E. (2018). Comparison of training approaches for photovoltaic forecasts by means of machine learning. *Applied Sciences (Switzerland)*, *8*(2). https://doi.org/10.3390/app8020228

Dolara, A., Leva, S., & Manzolini, G. (2015). Comparison of different physical models for PV power output prediction. *Solar Energy*, *119*, 83–99. https://doi.org/10.1016/j.solener.2015.06.017

Elango, B., & Rajendran, P. (2012). Authorship trends and collaboration pattern in the marine sciences literature: a scientometric study. *International Journal of Information Dissemination and Technology*, (January).

energysage. (2019). How solar panel cost and efficiency have changed over time. Recuperado el 2 de diciembre de 2020, de https://news.energysage.com/solar-panel-efficiency-cost-over-time/

Fernandes, C. A. F., Torres, J. P. N., Morgado, M., & Morgado, J. A. P. (2016). Aging of solar PV plants and mitigation of their consequences. *Proceedings - 2016 IEEE International Power Electronics and Motion Control Conference, PEMC 2016*, 1240–1247. https://doi.org/10.1109/EPEPEMC.2016.7752174

Fritts, C. E. (1883). On a New Form of Selenium Photocell. *American J. of Science*, *26*, 465.

Fuentes, M., Nofuentes, G., Aguilera, J., Talavera, D. L., & Castro, M. (2007). Application and validation of algebraic methods to predict the behaviour of crystalline silicon PV modules in Mediterranean climates. *Solar Energy*, *81*(11), 1396–1408. https://doi.org/10.1016/j.solener.2006.12.008

Geisz, J. F., France, R. M., Schulte, K. L., Steiner, M. A., Norman, A. G., Guthrey, H. L., … Moriarty, T. (2020). Six-junction III–V solar cells with 47.1% conversion efficiency under 143 Suns concentration. *Nature Energy*, *5*(4), 326–335. https://doi.org/10.1038/s41560-020-0598-5

GURU 99. (2020). Supervised vs Unsupervised Learning: Key Differences. Recuperado el 4 de mayo de 2020, de https://www.guru99.com/supervised-vs-unsupervised-learning.html

Hagan, M., & Demuth, H. (2014). Neural Network Design. En M. Hudson & O. De Jesús (Eds.), *Neural Networks in a Softcomputing Framework* (2a ed.). Recuperado de hagan.okstate.edu/nnd.html

Han, S., Qiao, Y. hui, Yan, J., Liu, Y. qian, Li, L., & Wang, Z. (2019). Mid-to-long term wind and photovoltaic power generation prediction based on copula function and long short term memory network. *Applied Energy*, *239*(January), 181–191. https://doi.org/10.1016/j.apenergy.2019.01.193

Haque, A. U., Nehrir, M. H., & Mandal, P. (2013). Solar PV power generation forecast using a hybrid intelligent approach. *2013 IEEE Power & Energy Society General Meeting*, (D), 1–5. https://doi.org/10.1109/PESMG.2013.6672634

Hartigan, J. A., & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Applied Statistics*, *28*(1), 100. https://doi.org/10.2307/2346830

Igel, C., & Hüsken, M. (2000). Improving the Rprop learning algorithm. *Proceedings of the Second*

International Symposium on Neural Computation, 115–121. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.17.3899&rep=rep1&type=pdf

Inman, R. H., Pedro, H. T. C., & Coimbra, C. F. M. (2013). Solar forecasting methods for renewable energy integration. *Progress in Energy and Combustion Science*, *39*(6), 535–576. https://doi.org/10.1016/j.pecs.2013.06.002

Isha, Chaudhary, A. S., & Chaturvedi, D. K. (2020). Effects of Activation Function and Input Function of ANN for Solar Power Forecasting. En *Lecture Notes in Networks and Systems* (Vol. 94, pp. 329–342). https://doi.org/10.1007/978-981-15-0694-9_31

Isik, F., Ozden, G., & Kuntalp, M. (2012). Importance of data preprocessing for neural networks modeling: The case of estimating the compaction parameters of soils. *Energy Education Science and Technology Part A: Energy Science and Research*, *29*(2), 871–882.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. https://doi.org/10.1007/978-1-4614-7138-7

Kassambara, A. (2017). *Machine Learning Essentials: Practical Guide in R* (1a ed.). Marseille: STHDA.

Keith, T. Z. (2019). Multiple regression and beyond: An introduction to multiple regression and structural equation modeling. En *Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling*. https://doi.org/10.4324/9781315162348

Kubat, M. (2017). An Introduction to Machine Learning. En *An Introduction to Machine Learning*. https://doi.org/10.1007/978-3-319-63913-0

Kudo, M., Takeuchi, A., Nozaki, Y., Endo, H., & Jiro, S. (2009). Forecasting electric power generation in a photovoltaic power system for an energy network. *Electrical Engineering in Japan (English translation of Denki Gakkai Ronbunshi)*, *167*(4), 16–23. https://doi.org/10.1002/eej.20755

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling* (1a ed.). https://doi.org/10.1007/978-1-4614-6849-3

Leva, S., Dolara, A., Grimaccia, F., Mussetta, M., & Ogliari, E. (2017). Analysis and validation of 24 hours ahead neural network forecasting of photovoltaic output power. *Mathematics and Computers in Simulation*, *131*, 88–100. https://doi.org/10.1016/j.matcom.2015.05.010

Li, Y., Su, Y., & Shu, L. (2014). An ARMAX model for forecasting the power output of a grid connected

photovoltaic system. *Renewable Energy*, *66*, 78–89. https://doi.org/10.1016/j.renene.2013.11.067

Liu, J., Fang, W., Zhang, X., & Yang, C. (2015). An Improved Photovoltaic Power Forecasting Model With the Assistance of Aerosol Index Data. *IEEE Transactions on Sustainable Energy*, *6*(2), 434–442. https://doi.org/10.1109/TSTE.2014.2381224

Liu, Z., Li, L., Tseng, M., & Lim, M. K. (2020). Prediction short-term photovoltaic power using improved chicken swarm optimizer - Extreme learning machine model. *Journal of Cleaner Production*, *248*(xxxx), 119272. https://doi.org/10.1016/j.jclepro.2019.119272

Loy, J. (2019). *Neural network projects with Python : the ultimate guide to using Python to explore the true power of neural networks through six projects* (1a ed.). Recuperado de https://books.google.com.co/books?id=6AuLDwAAQBAJ

Lubo, U. D. (2019). Cargos de respaldo por uso de la red eléctrica en el costo unitario de energía distribuida: desafíos y oportunidades para la planificación. *Revista UIS Ingenierías*, *18*(3), 67–74. https://doi.org/10.18273/revuin.v18n3-2019007

Marx, W., Bornmann, L., Barth, A., & Leydesdorff, L. (2014). Detecting the historical roots of research fields by reference publication year spectroscopy (RPYS). *Journal of the Association for Information Science and Technology*, *65*(4), 751–764. https://doi.org/10.1002/asi.23089

Matasci, S. (2018). How Solar Panel Cost and Efficiency have Changed Over Time. Recuperado el 12 de julio de 2019, de Energy Sage website: https://news.energysage.com/solar-panel-efficiency-cost-over-time/

MathWorks. (2020). Machine Learning vs Deep Learning. Recuperado el 18 de marzo de 2020, de https://explore.mathworks.com/machine-learning-vs-deep-learning/chapter-1-129M-100NU.html

Moor, J. (2006). Artificial Intelligence Conference: The Next Fifty Years. *AI Magazine*, *27*(4), 87–91. https://doi.org/https://doi.org/10.1609/aimag.v27i4.1911

Mosa, M., Shadmand, M. B., Balog, R. S., & Rub, H. A. (2017). Efficient maximum power point tracking using model predictive control for photovoltaic systems under dynamic weather condition. *IET Renewable Power Generation*, *11*(11), 1401–1409. https://doi.org/10.1049/iet-rpg.2017.0018

Park, N. C., Oh, W. W., & Kim, D. H. (2013). Effect of temperature and humidity on the degradation rate of multicrystalline silicon photovoltaic module. *International Journal of Photoenergy*, *2013*.

https://doi.org/10.1155/2013/925280

Park, N., Kim, J. H., Kim, H. A., & Moon, J. C. (2017). Development of an algebraic model that predicts the maximum power output of solar modules including their degradation. *Renewable Energy*, *113*, 141–147. https://doi.org/10.1016/j.renene.2017.05.073

Pedro, H. T. C., & Coimbra, C. F. M. (2012). Assessment of forecasting techniques for solar power production with no exogenous inputs. *Solar Energy*, *86*(7), 2017–2028. https://doi.org/10.1016/j.solener.2012.04.004

Pelland, S., Galanis, G., & Kallos, G. (2013). Solar and photovoltaic forecasting through post-processing of the Global Environmental Multiscale numerical weather prediction model. *Progress in Photovoltaics: Research and Applications*, *21*(3), 284–296. https://doi.org/10.1002/pip.1180

Ponce Cruz, P. (2010). Inteligencia Artificial con aplicaciones a la ingeniería. En *Alfaomega Grupo Editor, S.A.* (1a ed.). México D.F.

Pulipaka, S., & Kumar, R. (2016). Power prediction of soiled PV module with neural networks using hybrid data clustering and division techniques. *Solar Energy*, *133*, 485–500. https://doi.org/10.1016/j.solener.2016.04.004

PyRP. (2020). K-Means. Recuperado el 4 de mayo de 2020, de http://pypr.sourceforge.net/kmeans.html

Ramsami, P., & Oree, V. (2015). A hybrid method for forecasting the energy output of photovoltaic systems. *Energy Conversion and Management*, *95*, 406–413. https://doi.org/10.1016/j.enconman.2015.02.052

Rana, M., & Rahman, A. (2020). Multiple steps ahead solar photovoltaic power forecasting based on univariate machine learning models and data re-sampling. *Sustainable Energy, Grids and Networks*, *21*, 100286. https://doi.org/10.1016/j.segan.2019.100286

Raza, M. Q., & Khosravi, A. (2015). A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. *Renewable and Sustainable Energy Reviews*, *50*, 1352–1372. https://doi.org/10.1016/j.rser.2015.04.065

Rhaiem, M., & Bornmann, L. (2018). Reference Publication Year Spectroscopy (RPYS) with publications in the area of academic efficiency studies: what are the historical roots of this research topic? *Applied Economics*, *50*(13), 1442–1453. https://doi.org/10.1080/00036846.2017.1363865

Riedmiller, M., & Braun, H. (1993). Direct adaptive method for faster backpropagation learning: The RPROP

algorithm. *1993 IEEE International Conference on Neural Networks*, 586–591. https://doi.org/10.1109/icnn.1993.298623

Rodríguez-Benítez, F. J., Arbizu-Barrena, C., Huertas-Tato, J., Aler-Mur, R., Galván-León, I., & Pozo-Vázquez, D. (2020). A short-term solar radiation forecasting system for the Iberian Peninsula. Part 1: Models description and performance assessment. *Solar Energy*, *195*(June 2019), 396–412. https://doi.org/10.1016/j.solener.2019.11.028

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Skoplaki, E., & Palyvos, J. A. (2009). On the temperature dependence of photovoltaic module electrical performance: A review of efficiency/power correlations. *Solar Energy*, *83*(5), 614–624. https://doi.org/10.1016/j.solener.2008.10.008

Vikram, A. (2020). What are the most efficient solar panels on the market? Solar panel cell efficiency explained. Recuperado el 20 de mayo de 2020, de Energysage website: https://news.energysage.com/what-are-the-most-efficient-solar-panels-on-the-market/

Yang, H.-T., Huang, C.-M., Huang, Y.-C., & Pai, Y.-S. (2014). A Weather-Based Hybrid Method for 1-Day Ahead Hourly Forecasting of PV Power Output. *IEEE Transactions on Sustainable Energy*, *5*(3), 917–926. https://doi.org/10.1109/TSTE.2014.2313600

Zhai, Y. (2005). Time series forecasting competition among three sophisticated paradigms. University of North Carolina.

Zhang, Y., Chen, G. P., Malik, O. P., & Hope, G. S. (1993). An Artificial Neural Network Based Adaptive Power System Stabilizer. *IEEE Transactions on Energy Conversion*, *8*(1), 71–77. https://doi.org/10.1109/60.207408

Zhu, H., Li, X., Sun, Q., Nie, L., Yao, J., & Zhao, G. (2016). A power prediction method for photovoltaic power plant based on wavelet decomposition and artificial neural networks. *Energies*, *9*(1), 1–15. https://doi.org/10.3390/en9010011